SUPPLEMENTARY INFORMATION

The *Pleurobrachia bachei* Genome

Leonid L. Moroz, Kevin M. Kocot*, Mathew R. Citarella*, Sohn Dosung, Tigran P. Norekian, Inna S. Povolotskaya, Anastacia P. Grigorenko, Christopher Dailey, Eugene Berezikov, Kate Buckley, Denis Reshetov, Andrey Ptitsyn, Krishanu Mukherjee, Tatiana P. Moroz, Rachel Sanford, Emily Dabe, Fahong Yu, Vladimir V. Kapitonov, Jerzy Jurka, Yuri Bobkov, Kirill Ukhanov, Joshua J. Swore, David O. Girardo, Yelena Bobkova, Alexander Fodor, Nathan Churches, Gabrielle Winters, Caleb Botchwik, Rebecca Bruders, Fedor Gusev, Ellen Kittler, Claudia E. Mills, Jonathan Rast, Romain Derelle, Victor V. Solovyev, Fyodor A. Kondrashov, Billie J. Swalla, Jonathan V. Sweedler, Evgeny I. Rogaev, Kenneth M. Halanych, Andrea B. Kohn

*equal contribution

Supplementary Methods: Sections **SM1-SM17** (pp-4-36).

Supplementary Tables: 1S-30S (pp-37-67) including 4 Excel Tables 27-30S as separate items

Supplementary Data: Sections **SD1-SD5** (pp-68-104) and references (105-122).

Supplementary Methods **Table of Contents**

S2.2.2 SOAPdenovo Assembly
S2.2.3 Pseudo-454 Hybrid Assembly
S2.2.4 *Pleurobrachia bachei* Draft Genome Assembly v1.1
S2.2.5 Polymorphism
SM2.3 Prokaryotic Contamination Removal

SM3 *Pleurobrachia bachei* Genome Annotation
SM3.1 Gene Predictions
S3.1.1 Augustus Predictions
S3.1.2 Fgenesh and Fgenesh++ Predictions
S3.1.3 Extending Partially Predicted Genes
SM3.2 Annotation of *P. bachei* Gene Models
S3.2.1 Homology of Gene Models to Public Data Sets
S3.2.2 Identification of Transposable elements

SM4 *Pleurobrachia bachei* Gene Content
SM.4.1 Identification of *P.bachei* Genes and Gene Model Analysis
S4.1.1 Homology of Gene Models to Public Protein Data
S4.1.2 Gene Ontology Terms
S4.1.3 KEGG Pathways
S4.1.4 UniProtKB/Swiss-Prot to KEGG Associations
S4.1.5 Pfam Domains
SM4.2 Transcriptome Analysis
S4.2.1 Reference Transcriptome Assembly
S4.2.1.1 454 Assembly
S4.2.1.2 Illumina Assembly
S.4.2.1.3 Ion Torrent Assembly
S4.2.2 Reference Transcriptome Annotation
S4.2.3 RNA Expression Data
S4.2.3.1 Gene Model Expression Levels Across Transcriptomes
S4.2.3.2 454 Transcriptome Expression
S4.2.3.3 Illumina Transcriptome Expression
S4.2.3.4 Ion Torrent Transcriptome Expression
S4.2.3.5 Identification of *P. bachei* Orthologs
S4.2.3.6 RNA-seq Expression Profiling
S4.2.3.7 Prediction of Secretory Peptides

SM5 Data Availability and Visualization
S5.1 Genome Project Homepage
S5.2 UCSC Genome Browser Mirror
S5.3 Comparative Transcriptomics Database
S5.4 Direct Download

SM6 Primary Accessions

SM7 Phylogenomic Analyses

- S7.1 Methods for Phylogenomic Analyses
- S7.2 Analyses with Expanded Ctenophora sampling
- S7.2 Analysis of individual genes and families
- S7.3 Analysis of gene gain and gene loss
- S7.4 Taking into account protein divergence for phylogeny inference

SM8 Analysis of DNA methylation

SM9 Molecular Cloning

- S9.1 cDNA library construction
- S9.2 Cloning

SM10 *In situ* Hybridization Protocol

SM11 Immunohistochemistry

SM12 Scanning Electron Microscopy

SM13 Electrophysiological Methods

SM14 Calcium Imaging

SM15 Solutions and chemicals

SM16 Pharmacological and Behavioral Assays

SM17 Determination of the Presence of Classical Neurotransmitters by Capillary Electrophoresis (Metabolomics/Microchemical Assays)

SM1 *Pleurobrachia bachei* Genome and Transcriptome Sequencing

SM1.1 **Animal Collections**

Pleurobrachia bachei and most other ctenophore species (*Euplokamis dunlapae, Dryodora glandiformis, Beroe abyssicola, Mnemiopsis leidyi , Bolinopsis infundibulum* and *Mertensiidae sp.*), were randomly collected from the dock at Friday Harbor Laboratories (University of Washington) near Seattle (Pacific North-Western Coast of USA) during 2007-2013. Animals were maintained in running Pacific sea water for up to two weeks and used for both the genome sequencing as well as all reported experiments including RNA-seq, *in situ* hybridization, mapping, metabolomic, developmental and physiological studies. Additional ctenophore species were collected at the Atlantic coast of Florida and around of Woods Hole, Massachusetts (*Pleurobrachia pileus, Pleurobrachia sp*., *Mnemiopsis leidyi*) as well as central Pacific (Palau, Hawaii, *Coeloplana astericola, Vallicula multiformis*); they were maintained in local waters for up to 1-2 days before experiments.

Before DNA extractions for genomic analysis, all animals were microscopically examined for potential ectoparasites such as copepods, and then washed three times for at least 30 minutes in twice filtered (0.2 µm) sea water. Animals were anesthetized in 60% (volume/body weight) isotonic $MgCl_2$ (337mM). Specific tissues were surgically removed with sterile fine forceps and scissors, again washed at least twice in filtered sea water, and processed for RNA isolations. Whole animals, sometimes partially dissected were processed for genomic DNA isolations.

SM1.2 **Genomic DNA isolation**

DNA was immediately isolated with Genomic-tip 100/G and Genomic DNA Buffer Set (Cat # 10243, 19060, QIAGEN Inc.,Valencia, CA) according to manufacturer's recommendations. Animals were placed in lysis buffer G2 with RNase A and proteinase K, then incubated at 50º C until the sample just dissolved. All samples were vortexed and passed over the Genomic-tip 100/G column. Genomic DNA was eluted in 150 µL of MilliQ water. Quality and quantity of genomic DNA was analyzed on a Qubit® 2.0 Fluorometer (Life technologies) and an agarose gel.

SM1.3 **RNA isolation**

Selected tissues were fixed in 75% ethanol for several minutes and RNA was isolated immediately. Total RNA was extracted from small amounts of tissue using RNAqueous-Micro® (Cat # 1912, Applied Biosystems/Ambion/Life Technology, Austin, TX) or RNAqueous® (Cat # 1931, Applied Biosystems/Ambion/ Life Technology, Austin, TX) for large amounts of tissue. Before further processing of RNA, quality was checked using a 2100 Bioanalyzer™ (Agilent Technologies, Santa Clara, CA). Extracted RNA was treated with 2 µL of DNAase (2 Units/ul) for 15 min at 37° C (TURBO DNA-free™ Kit Cat # AM1907, Applied Biosystems/Ambion/ Life Technology, Austin, TX). The sample was ready for library construction.

SM1.4 Genomic DNA Sequencing Library Construction

SM1.4.1 454/Roche Genomic DNA libraries

All 454/Roche genomic DNA libraries were constructed at the University of Florida core facilities (ICBR) or SeqWright, Inc., Houston, TX. For the shotgun library 1 µg of genomic DNA was sheared and ligated with 454 adaptors. Paired end libraries were constructed from a pool of 10 animals or 20 µg of genomic DNA. Genomic DNA was fragmented to appropriate size and ligated with biotinylated hairpin linker adaptor. DNA was exonuclease treated, then adaptors were cleaved and the DNA circularized. Circularized DNA was fragmented and enriched by streptavidin capture of biotinylated adaptors. Sequencing adaptors were finally ligated onto the fragmented enriched DNA and sequenced. Sequencing was performed on titanium 454 plates. All genomic sequences were submitted to NCBI on SRA accession number Project: SRP001155.

SM1.4.2 Illumina Genomic DNA Libraries and Sequencing

The paired end library was made from one animal with the Illumina Paired-End Kit according to manufacturer's recommendations (Cat # PE-930-1001, Illumina Inc., San Diego, CA). 1 µg of genomic DNA, isolated as described in S1.2, was fragmented using a Bioruptor® sonication device (Diagenode Inc., Denville, NJ) to recommended size. Ends of fragmented DNA were repaired and ligated with Illumina paired-end adaptors. Products were purified, amplified and sequenced. Sequencing was performed on both a GA machine and a HiSeq 2000 Illumina sequencing machine. Mate pair libraries were prepared with 2–5kb samples using a Mate Pair Library kit (Cat # 1005363 Rev. B, Illumina Inc., San Diego, CA). 10 µg of genomic DNA from a pool of about dozen animals was sheared to the appropriate size on a HydroShear (Digilab Inc., Holliston, MA). Ends of genomic DNA were repaired, biotinylated and circularized. Circularized DNA was then fragmented and enriched for biotinylated fragments. These ends were repaired and Illumina sequencing adaptors ligated. This product was amplified and purified. Sequencing was performed on both a GA machine and a HiSeq2000 Illumina sequencing machine. All genomic sequences were submitted to NCBI on SRA accession number Project: SRP001155. A summary of the genomic libraries constructed is in **Supplementary Table 1S**. All genomic sequencing is summarized in **Supplementary Table 2S**.

The MiSeq Illumina paired end libraries were constructed from genomic DNA isolated as described in S1.2 from a single animal. Then 1 µg of gDNA was fragmented to 800bp with the Covarus Focused-ultrasonicator M220 (Covarus, Woburn, MA). The fragmented gDNA was end repaired, A tailed, ligated and amplified using the NEBNext® Fast DNA Library Prep Set for Illumina (Cat. # E6040S, New England Biolabs). The final product was sized selected to 738 bp and purified with Agencourt AMPure XP beads (Cat. # A63880, Beckman Courlter). Concentration and size was determined with the 2200 TapeStation Instrument (Cat. # G2964AA, Agilent Technologies). Finally a 2 x 250bp sequencing was performed on a MiSeq Illumina sequencing machine. All genomic sequences were submitted to NCBI on SRA accession number Project: SRP001155. A summary of the genomic libraries constructed is in **Supplementary Table 1S**. All genomic sequencing is summarized in **Supplementary Table 2S**.

SM1.5 **Transcriptome Library Construction**

A transcriptome is comprised of a subpopulation of RNA molecules derived from those genes whose biological information is required by a cell at a particular time. To characterize a transcriptome (RNA-seq), every RNA molecule needs to be identified and relative abundances determined; high through-put sequencing is the ideal technology for this approach. We used three sequencing technology platforms: 454 Life Sciences platform (454 Life Sciences, a Roche company, Branford, CT), Illumina platform (Illumina Inc., San Diego, CA), and Ion Torrent PGM (Life Technologies, Carlsbad, CA). Specifically, we utilized the long reads of the 454/Roche platform, the high through-put of Illumina and the speed of Ion Torrent (Life Technologies).

We developed a reduced representation sequencing protocol for the 454 and Ion Torrent sequencing platforms that can detect low abundance transcripts[1]. The method greatly reduces the amount of sequencing and gives more accurate quantification. Basically, sequential addition of adaptors paired with controlled digestion of cDNA generates a one transcript per one read type of analysis. We show that our sequencing libraries maintain the directional context, thus allowing analysis of sense and antisense transcripts. All additional details of the procedure are reported elsewhere[1,2].

SM1.5.1 **454/Roche Transcriptome Libraries**

Starting quantities of RNA ranged from 20 ng to 2 µg. Initially, we tested RiboMinus™ Eukaryote Kit (Cat # A10837-08, Invitrogen, Life Technologies) but this procedure did not efficiently remove all rRNA from ctenophores; so total RNA was used for these libraries. For the large 454 FLX projects one animal was used and the tissue processed. Amplified cDNA was generated from total RNA using the Marathon cDNA amplification kit (Cat # 634913, Clontech Laboratories, Inc., Mountain View, CA). The first strand synthesis utilized the AMV Reverse Transcriptase and an oligo(dT) primer, Trsa. After second strand synthesis following the Marathon kit protocol, the cDNA was purified with AMPure XP Reagent (Cat # A63880, Agencourt, Beckman Coulter, Inc). cDNA was ligated with the double stranded A adaptor at a final concentration of 1 µM along with T4 DNA ligase from the Quick Ligation™ Kit (Cat# M2200S, New England BioLabs) for 10 minutes at room temperature. cDNA was fractionated by digestion with 20 units of Alu 1 and NEBbuffer 2 (Cat # RO137S0, New England BioLabs, Ipswich, MA) for 1 hour at 37° C. A second double stranded adaptor, B was ligated at a final concentration of 1 µM with the digested cDNA using the Quick Ligation™ Kit (Cat # M2200S, New England BioLabs) for 10 minutes at room temperature. cDNA was purified using AMPure XP Reagent Preparation according to the manufacturer's recommendations and eluted in 16 µL of RNAase-free water. Amplification of cDNA was performed using 16 µL of the purified cDNA, 10× LA PCR Buffer II, dNTP and TaKaRa LA Taq™ (5 units/µl) (Cat # TAK RR002A, TaKaRa, Madison, WI). The A pcr and B pcr primers were added to the amplification mix at a final concentration of 200 nM. The PCR amplification protocol consisted of 17 cycles of 95°C for 30 sec, 60°C for 30 sec, 72°C for 1 min. PCR products were purified using AMPure XP Reagent Preparation according to the manufacturer's recommendations. Sequencing of amplicon products used the GS FLX emPCR Kit II (Amplicon A) (Cat # 25740110, Roche Applied Science) and GS FLX Sequencing Kit (70x75) (Cat # 05233526001, Roche Applied Science).

Sequencing was performed at the University of Florida core facilities (ICBR) and SeqWright, Inc., Houston, TX. All additional details of the procedure are reported elsewhere[1]. All transcriptome sequences were submitted to NCBI on SRA accession number Project: SRP000992.

SM1.5.2 **Illumina Transcriptome Libraries**

Two Illumina RNA-seq libraries were constructed; one total RNA library was isolated from one adult animal and a second from a mixture of embryos for *P. bachei*. For both libraries, the mRNA Sequencing Sample Preparation kit was used according to the manufacturer's recommendations (Cat # RS-930-1001, Illumina Inc., San Diego, CA). Briefly, mRNA was isolated from 1 µg of total RNA on Sera-Mag Magnetic Oligo(dT) Beads. Purified mRNA was fragmented and reverse transcribed to make cDNA. The second strand was made and 3' ends were adenylated, then Illumina adaptors were ligated to this fragmented dsDNA. Fragmented dsDNA was gel purified to the appropriate size faction, then amplified by PCR. PCR products were purified and sequenced on an Illumina GA II Sequencing System and a HiSeq2000 Illumina Sequencing System. All transcriptome information including tissues used for construction of libraries and sequencing is summarized in **Supplementary Table 3S** and all sequence data was submitted to NCBI on SRA accession number Project: SRP000992.

Illumina libraries for species of ctenophores other than *P. bachei* were constructed using a template-switch method of first and second strand synthesis using SMART™ cDNA Library Construction Kit (Clontech Cat # 634901). Full-length cDNA was then amplified using the Advantage 2 PCR system (Clontech) using the minimum number of PCR cycles necessary and sent to the HudsonAlpha Institute for Biotechnology (Huntsville, AL, USA) for shearing and 2 X 100 bp paired-end library preparation and Illumina sequencing. Each library was sequenced using approximately one-sixth of an Illumina HiSeq 2000 lane. Additional details of the procedure are reported elsewhere[1]. All transcriptome information including species used for construction of libraries and sequencing is summarized in **Supplementary Table 3S** and all sequence data was submitted to NCBI on SRA accession number Project: SRP000992.

SM1.5.3 **Ion Torrent RNA-seq Libraries**

RNA-seq libraries were constructed for all developmental stages of *P. bachei* from 1 cell through to 3 Day embryos as reported previously[1].

All transcriptome information including development stages used for construction of libraries and sequencing is summarized in **Supplementary Table 3S** and all sequence data was submitted to NCBI on SRA accession number Project: SRP000992.

SM2 Sequencing and Assembly of the *Pleurobrachia bachei* Genome

SM.2.1 **Genome Sequencing**

SM2.1.1 **Shotgun Sequencing**

Shotgun sequencing libraries were prepared for the 454/Roche and Illumina sequencing platforms as described in section S1.4.2. 3,286,344 single-end reads and 2,145,046 paired-end reads were sequenced on the 454 FLX Titanium sequencing platform. These data total 1,177,702,794 base-pairs (bp) for the single-end sequence reads and 746,499,725 bp for the paired-end sequence reads. Additionally, 540,043,926 paired-end sequence reads totaling 71,058,073,560 bp were sequenced using the Illumina HiSeq 2000, Ilumina GAIIx and Illumina MiSeq sequencing platforms. All of these data sets are available at NCBI's SRA, accession # Project: SRP001155. The results of the shotgun sequencing are summarized in **Supplementary Table 2S**.

SM2.1.2 **Mate-Pair Sequencing**

Mate-pair sequencing libraries were prepared for the Illumina GAIIx and HiSeq 2000 sequencing platforms as described in section S1.4.2. A total of 59,033,324,028 bp of sequence data in 612,457,003 sequence reads were obtained from five combined runs. Mate pair libraries feature insert sizes in the range of 180-530bp, and summary statistics for mate-pair sequencing can be found in **Supplementary Table 2S**. All mate-pair sequencing results are available on NCBI's SRA site, accession # Project: SRP001155.

SM2.2 ***De Novo* Genome Assembly Strategy**

The *Pleurobrachia bachei* draft genome was assembled using a custom approach designed to leverage the individual strengths of three popular *de novo* assembly packages and strategies: Velvet[3] (described in S2.2.1), SOAPdenovo[4] (described in S2.2.2), and pseudo-454 hybrid assembly with ABySS[5] (described in S2.2.3). Input data sets for each *de novo* assembly were the 454 and Illumina reads (**Supplementary Tables 1S-3S**). Contribution of non-redundant scaffolds above 2kb by each of the individual *de novo* assemblies to the *P. bachei* draft genome version 1.0 was decided as described in S2.2.4.

Briefly, genes were predicted in each of the three sets of genomic scaffolds using Augustus version 2.5.5[6], initially trained on *Amphimedon queenslandica*, then using several hundred *Pleurorachia bachei* genes from RNA-seq profiling and about one hundred full-length genes manually cloned over the course of the project in the Moroz laboratory. The three obtained sets of gene models were then aligned against themselves with blastp to remove intra-assembly redundancy. The resulting non-redundant lists were then pairwise-aligned to each other to remove inter-assembly redundancy. Scaffolds from each assembly that contributed a non-redundant gene model were included in the initial draft assembly of the genome (21,982 scaffolds containing 156,146,497 bp). Draft version 1.1 of the genome was produced by screening and removing contamination from prokaryotic sequences as described in the S2.3. Full summary statistics for the draft genome of *P. bachei* are presented in **Supplementary Table 5S**.

SM2.2.1 **Velvet Assembly**

The Velvet *de novo* assembly contributed 17,494 scaffolds to draft assembly version 1.1 of the *Pleurobrachia bachei* genome.

Assembly of *Illumina* reads was done in three steps. First, a short insertion length library (mean 320 bp) was assembled using Velvet software version 1.0.18 into 637,735 contigs achieving a total length of 153 Mbp and N50 of 2,476. Prior to the assembly, reads containing ambiguous characters were removed and B-tails of the reads were trimmed, resulting in 259,174,104 high-quality reads left with minimal length of 52. The second step included scaffolding of the contigs using SSPACE software version 1.1. Paired-end information was used from the shortest (320 bp) insertion length library to the largest one (4-5 kbp) with the resulting assembly achieving total length of ~170 Mbp and scaffold N50 of 16,518bp.

SM2.2.2 **SOAP *de novo* Assembly**

We used Release 1.05 of SOAPdenovo software (part of the Short Oligonucleotide Analysis Package[4], and compiled the software locally from the source code on 02/14/2011. Several combinations of the basic input parameters have been tested to balance maximal inclusion of fragment libraries from different platforms and different experimental designs with maximal performance using our available computational resources. This testing resulted in use of a k-mer parameter set to 56 (application SOAPdenovo-63mer) for a balance between shorter k-mer settings for 454 libraries and longer settings for *Illumina* libraries. In this assembly, the Illumina GAIIx/HiSeq2000 reads were used as input during contig building and scaffolding of the initial assembly. 454 data was used exclusively during the scaffolding phase of the assembly process. The finished de novo assembly produced by this software package contained 258,775,366 bp in 1,020,687 assembled sequences with an n50 of 280bp and n90 of 130bp. This *de novo* assembly contributed 2,149 scaffolds to draft assembly version 1.1 of the *Pleurobrachia bachei* genome.

To verify that our draft hybrid *Pleurobrachia* assembly fully captures the genetic content present in the SOAPdenovo assembly (i.e. we did not throw information away during the assembly merge process), we aligned the SOAPdenovo assembly to the draft genome with bowtie2[7] using the very-sensitive-local setting. The alignment resulted in 97.84% of the SOAPdenovo assembly aligning to the draft genome, with 84.46% aligning to more than one position, lending support to the notion that many of the small contigs contained in the SOAPdenovo assembly are poorly-assembled repetitive regions. While some of these contigs may be the result of haplotype splitting during the assembly process, our analyses suggest that the main reason for the larger size of the SOAPdenovo assembly is that the SOAPdenovo assembler did a poor job of assembling the *Pleurobrachia bachei* genome.

SM2.2.3 **Pseudo-454 Hybrid Assembly**

To generate a *de novo* genome sequence for a eukaryotic genome by combining sequence information from different technologies, we used a hybrid approach for sequencing and

assembly, including the ABySS version 1.2.6 (Assembly by Short Sequences)[5] and Newbler version 2.3 (454 Life Science, Branford, CT) assemblers. Genomic data include the SE (single-end) and PE (paired-end) reads acquired from 454 and Illumina HiSeq/GAIIx sequencing technologies described in S1.1.4.

In data preparation with the programs of cross_match in Phrap and the cleanup module in PTA (Paracel Transcript Assembler) version 3.0.0 (Paracel Inc, Pasadena, CA), a high-quality region of every read was determined with respect to a quality value cutoff of 20 (phred-like score), and low-quality end regions before and after the high-quality region were trimmed. *Enterobacteria* phage contamination in the HiSeq data set was identified and removed from input sequences. Resulting reads with >=30 bp were kept for assembly. In ABySS, the assembly was performed in two major steps. First, without using paired-end information, the k-mer data set was processed to remove read errors and initial contigs are built. In the second step, paired-end information is used to extend contigs by resolving ambiguities in contig overlaps. Given the promising initial assembly of the Illumina PE read data, we assessed trimming and filtering as a means to improve the ABySS assembly accuracy. To find the optimal value of k, we ran multiple trials (a series of k-mers from 30 bp to 75 bp) and inspected the results.

To optimize the *de novo* genome assembly, assembled contigs with >=1,000 bp from ABySS were digitally sheared into 500 bp fragments, with 200 bp overlapping between two adjacent fragments. These fragments were then used as input to the Newbler assembler, and assembled with single- and paired-end reads from 454 sequencing. All sequences were checked and masked for universal and species-specific vector sequences, adaptors, and PCR primers used in cDNA libraries. If reads were at least 150bp in length, they were retained. The final assembly contained of 29,253,241bp in 5,906 scaffolds, with a N50 of 5,773bp and an N90 of 2,585bp. This *de novo* assembly contributed 2,339 scaffolds to draft assembly version 1.1 of the *Pleurobrachia bachei* genome.

SM2.2.4 ***Pleurobrachia bachei*** **Draft Genome Assembly v1.1 (Supplementary Table 5S)**

The *Pleurobrachia bachei* draft genome assembly v1.1 is the result of our efforts to merge our three individual assemblies (SOAPdenovo, Velvet, and ABySS/Newbler as described in S2.2) into a minimal set of scaffolds that completely and accurately represents all genomic loci.

To finalize this minimal set of scaffolds, we first predicted and validated genes for each of the individual assemblies (Velvet, SOAPdenovo, and ABySS/Newbler) via the methods discussed in S2.1. Initial predictions resulted in 5,963 gene models predicted from the ABySS/Newbler assembly, 7,146 gene models predicted from the SOAPdenovo assembly, and 27,572 gene models predicted from the Velvet assembly. We then screened each set of gene models for internal redundancy with the blastp program from NCBI's BLAST+ software suite[8]. An alignment with blastp between two models was sufficient to categorize the query gene as redundant if: 1) the query gene model had 90% identity to the hit model, 2) the alignment between the two models had a bit score of at least 100, and 3) the query gene model was shorter than the hit gene model. From the original sets of gene models, the following numbers of models

from each assembly were deemed non-redundant: ABySS/Newbler – 5,867(98%); SOAPdenovo – 5,963(83%); Velvet – 22,368 (81%). These are the unfiltered gene model set **Supplementary Excel Table 29S**.

These three sets of non-redundant gene models were then pairwise aligned to each other with blastp, to remove redundant genes across assemblies. Again, an alignment with blastp between two models was sufficient to categorize the query gene as redundant if: 1) the query gene model had 90% identity to the hit model, 2) the alignment between the two models had a bit score of at least 100, and 3) the query gene model was shorter than the hit gene model. All non-redundant models from the three constituent assemblies were then combined. Our analysis resulted in a final set of 22,368 non-redundant gene models (2,339 were a result of the ABySS/Newbler assembly, 2,149 a result of the SOAPdenovo assembly, and 17,880 a result of the Velvet assembly). Scaffolds producing these gene models were pooled and then screened for prokaryotic contamination as described in S2.3 to produce the draft assembly version 1.0. Short scaffolds (<400bp) were then removed. This set of data is the filtered gene models **Supplementary Excel Table 30S**.The scaffolds present after these screening steps represent the *Pleurobrachia bachei* draft genome assembly version 1.1, the statistics for which can be found in **Supplementary Table 5S**.

We assesed our draft assembly's coverage of the *Pleurobrachia bachei* genome using two approaches. First, RNA-seq data from our 454, Ion Torrent, and Illumina sequencing projects described in **Supplementary Table 4S** were aligned to the draft assembly with bowtie2. Of these data, 94.28% of the 1,153,937 **454** reads, 90.29% of the 12,114,377 **Ion Torrent** reads, and 90.47% of 423,888,856 the **Illumina** reads aligned to the draft genome. To compliment this method, 138,900,000 paired-end genomic Illumina reads (250bpx2 or 34.7 Gb) from a single individual were also mapped to the final draft assembly. 88% of these reads mapped, suggesting that 12% could be an upper bound on the amount of genomic sequence missing from the initial draft assembly. This is undoubtedly an overestimate, as it is improbable that all of the unmapped reads are unique. To validate the accuracy of this claim, we performed additional downstream analyses using the Khmer package[9] and Blat[10]. First, we removed redundant reads from the set of 16,668,000 that failed to map to our draft assembly with Khmer's normalize-by-median.py script (options: –N 4 –x 100e04 –C 5 –k 32 and an estimated false positive rate of 0.00). This step marked 67% of the unmapped reads as redundant, i.e. having a median k-mer coverage of 5 or greater, and removed those reads from the set. The remaining 5,840,406 unmapped reads were then re-aligned to the *Pleurobrachia bachei* draft genome with the Blat aligner, using the built-in settings for DNA remapping. An additional 1,226,576 previously unmapped reads were aligned by Blat, leaving only ~3% of the reads from the original data set unmapped against the reference genome.

Given the above mapping rates, we are confident that the draft genome presented in this study suitably covers the genome of *Pleurobrachia bachei*.

SM2.2.5 **Polymorphism estimates**

To assess polymorphism we used both a targeted and a whole genome approach similar to procedures described in the recent *Amphimedon* (sponge) and *Hydra* genome papers[11,12].

Briefly, in the whole genome approach, genomic DNA sequencing data from Illumina MiSeq were aligned to all scaffolds using bowtie2[7]. Genomic variants (SNPs) were called using samtools's mpileup algorithm[13] . The resulting Variant Call Format (VCF) file was trimmed to reduce over-sequencing and under-sequencing bias by first calculating the median read depth at each SNP position, and then excluding those SNPs that were either one median absolute deviation (MAD) below or two MADs above the median sequence depth. Our reference genome was then split into 500bp windows, and the number of SNPs in each window was computed. The resulting SNP count distribution was well fit by a geometric model with a maximum likelihood estimate for *p* of 0.132. Taking the expected value of this distribution using the standard formula (1-*p/p*), we estimate that one SNP occurs approximately every 76bp in the *Pleurobrachia bachei* genome, yielding a global estimated polymorphism rate of 1.3%.

For the targeted approach, we used a single individual animal following isolation of genomic DNA and total RNA for a cDNA library constraction. Specific primers were designed for a representative gene set. First, to validate the gene models produced from the genome assembly, we PCR amplified regions of a representative set of eleven gene models (Wnt 1, NB#3469705; Wnt2a, NB# 3467531; Innexin4, NB# 3464407; P2X, NB# GU395551; Synaptobrevin, NB# 3474909; BMP3, NB# 3467338; BMP4, NB# 3466704; Tubulins, NB# 3461972, NP1, NB#3479560, NB# 3462463; ELAV2, NB# 3466232). The respective gene products were amplified, cloned and sequenced. In addition, we analyzed five noncoding regions of the *Pleurobrachia* genome for the same gene set. The amplified regions were evaluated with a minimum number of five sequence reads. For the coding genes the polymorphism can be in the range of up to 6.5% with an average of 2.5%. For the noncoding regions the polymorphism ranged from 0.3% to 3.8%.

SM2.3 **Prokaryotic Contamination Removal**

To determine levels of prokaryotic contamination in draft assembly scaffolds, a custom contaminants database was obtained from the National Center for Biotechnology Information's GenBank, RefSeq, TPA and PDB databases by downloading all nucleotide sequences assigned a TaxonID from either Archaea or Bacteria. Redundant entries in this contaminants database were removed by performing a BLASTN of the database against itself and removing sequences having greater than 90 percent identity to another sequence, with a minimum hit length of 90 percent of the query sequence's length, via a custom Perl script. In each instance of a hit, the longer of the hit or query sequence was chosen to remain in the contaminants database. Post-redundancy removal, the contaminants database contained 12,128,236,262 bases in 1,265,937 sequences from Archaea and Bacteria.

The resultant contaminants database was then aligned to the *Pleurobrachia bachei* draft assembly using UCSC's BLAT software package[10] (version 34, compiled 03/04/2011). Scaffolds matching any sequence in the contaminants database with greater than 90 percent identity over 90 percent of either the scaffold or contaminant were marked as a candidate for removal. Alignments involving these candidates were then manually examined before classifying them as contaminated scaffolds. Our screening resulted in the identification of six scaffolds containing contamination from Bacteria or Archaea, constituting less than .001% of the total bases in the assembly (alignments between scaffolds and the contaminant database totaled 4,837bp).

SM3 *Pleurobrachia bachei* Genome Annotation

SM3.1 Gene Model Prediction

SM3.1.1 Augustus Workflow

Models for protein-coding genes were predicted out of genomic scaffolds using the Augustus software package (version 2.5.5 http://augustus.gobics.de/,[6]), initially trained on *Amphimedon queenslandica*. We supplemented this training using several hundred *P. bachei* genes from RNA-seq profiling and about a hundred full-length genes cloned over the course of the project in the Moroz laboratory. Transcriptional/RNA-seq confirmation for predicted genes was provided to Augustus in the form of hint files, produced by aligning assembled 454 and Illumina transcriptomes to genomic scaffolds with BLAT (version 34, default parameters) and submitting the PSL alignment files to blat2hints.pl, a utility script provided in the Augustus package. To generate a preliminary set of protein-coding genes, each set of assembly scaffolds (SOAPdenovo, Velvet, and ABySS/Newbler) was individually submitted to Augustus with its respective hints file. These initial predictions included 5,963 gene models predicted from theABySS/Newbler assembly, 7,146 gene models predicted from the SOAPdenovo assembly, and 27,572 gene models predicted from the Velvet assembly. Statistics of genes predicted from each set of scaffolds are summarized in **Supplementary Table 7S**.

Individual sets of gene predictions were then merged. Each set of predicted gene models from the three assemblies was aligned to itself with NCBI's blastp program to remove internal redundancy. Gene models were considered redundant if they had at least 90% identity to another gene model in the same set (over the entire length of the two gene models) and the bit score of the alignment exceeded 100. The result of this step was that the number of gene models for each assembly was reduced to 5,867 for SOAPdenovo, 5,963for ABySS/Newbler, and 22,368 for Velvet. To remove inter-assembly redundancy, each non-redundant list of gene predictions was pairwise aligned to each other, again removing predicted models sharing over 90% identity and an alignment bit score over 100. This step resulted in the inclusion of 2,339 models from the ABySS/Newbler assembly, 2,149 from the SOAPdenovo assembly, and 17,880 from the Velvet assembly

Non-redundant gene models were aligned to our assembled 454 and Illumina transcriptomes with tblastn. 18,994 (85%) gene models were supported by the 454 transcriptome assembly and 15,954 (71%) were supported by the Illumina transcriptome assembly. In this instance, support of a gene model is considered to be at least one alignment of an assembled transcript to the model with greater than 90% identity over the length of the alignment. Taking into account the union of the two alignment results, a total of **19,996** (89%) of the predicted gene models obtained some level of support. Due to our extensive transcriptome sequencing (~1,790x coverage), we removed any predicted gene model lacking such support. An additional 492 partial gene models were merged with other models, as described in S3.1.3. This brought the final count of *Pleurobrachia* gene models to 19,523, **Supplementary Table 30S**

SM3.1.2 Fgenesh and Fgenesh++ Gene Predictions

Two gene prediction sets were generated at Softberry Inc using Fgenesh and the Fgenesh++ pipeline[14,15]. Fgenesh is a hidden Markov model based on an *ab initio* gene prediction program. Fgenesh++ is a pipeline for automatic prediction of genes, which, in addition to Fgenesh, includes sequence analysis software to incorporate information from full-length cDNA alignments and similar proteins from the eukaryotic section of the NCBI NR database[16]. Both Fgenesh and Fgenesh++ used sponge and bovine-specific gene-finding parameters developed specifically for *Pleurobrachia*.

SM3.1.3 Extending Partially Predicted Genes

A custom Java tool has been developed to merge and polish partially predicted, non-redundant gene models with assembled transcriptome data. The software attempts to extend partial gene model predictions based on evidence from transcriptome sequence by using the transcript to bridge 5' and 3' fragments of partially predicted genes. The software begins by aligning the non-redundant set of gene models to an assembled transcriptome data set with blastx from NCBI's BLAST+ software. A "best alignment" is chosen from each set of alignments of a given transcript to the gene models, maximizing the length of the alignment and the percent identity. Using this alignment as a seed, overlaps are called between the transcript and the gene model by examining the region 5' and 3' to the seed. If sequence similarity between the transcript and the gene model (excluding overhangs) is >= 90%, an overlap is called. After the maximal overlap for each gene model is calculated, the models are examined for extensions on their termini. Gene models are then merged with extending transcripts into a growing sequence chain. Additional chains or gene models are then added to existing chains based on overlaps. When every overlap in the dataset has been examined, the fully extended chain is printed as a merged gene model. An analysis of alignments of the non-redundant gene models to our assembled transcriptomes with this Java tool reduced the number of predicted genes by 492 to 19,523. Comparison of the *P. bachei* draft genome with other assembled genomes is found in **Supplementary Table 8S.** 19,523 of *Pleurobrachia* predicted genes were also used to identify their possible homologs in assembled transcriptomes from 10 other ctenophore species as summarized in the **Supplementary Tables 10S** and **11S.**

SM3.2 Annotation of *Pleurobrachia*'s Gene Models

SM3.2.1 Homology of Gene Models to Public Data Sets

Gene models were uploaded to the In-VIGO BLAST interface (https://invigo.acis.ufl.edu/icbr/invigo/Login), which was designed and hosted by the Advanced Computing and Information Systems Laboratory at the University of Florida. A blastp alignment of the gene models was performed against the entirety of NCBI's non-redundant protein database (http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastdb.html) and the Swiss-Prot protein database (http://www.uniprot.org/). A summary of the homology of *Pleurobrachia bachei* gene models to known protein coding genes from these public data sets at various levels of significance can be found in **Supplementary Table 7S**.

SM3.2.2 **Identification and Features of Transposable Elements (TE) in the *Pleurobrachia bachei* Genome**

Transposable elements were identified using WU-BLAST (http://blast.wustl.edu) and its implementation in CENSOR (http://girinst.org/censor/). First, we detected DNA sequences coding for proteins similar to transposases, reverse transcriptases, and DNA polymerases representing all known classes, superfamilies and clades of TEs described in the literature and/or collected in Repbase[17]. The detected sequences have been clustered based on their pairwise identities by using BLASTclust (standalone NCBI BLAST). Each cluster has been treated as a potential family of TEs described by its consensus sequence. Consensus sequences were built automatically based on multiple alignments of cluster sequences expanded in both directions and manually modified based on structural characteristics of known TEs. All autonomous non-LTR retrotransposons have been classified based on RTclass1[18]. See **Supplementary Table 9S** for summary of TEs in *P. bachei*.

SM4 *Pleurobrachia bachei* Gene Content

SM4.1 **Identification of *Pleurobrachia bachei* Genes and Gene Model Analysis**

SM4.1.1 **Homology of Gene Models to Public Protein Data Sets**

P. bachei gene models described in section S.3.1 were aligned to the National Center of Biotechnology Information's (NCBI) non-redundant protein database (NR), as well as the UniProtKB/*Swiss-Prot* (SP) protein database on the University of Florida's High Performance Computing Cluster (UFHPC). Alignment of all gene models was performed using blastp (parameters: parameters -evalue 10 -num_alignments 5 -num_descriptions 5) from NCBI's BLAST+ software package, keeping the five best alignments between each gene model and sequences from each protein database. Blastp jobs were submitted to UFHPC via the Moroz laboratory's zero-click automatic analysis pipeline[19], which was also used to monitor job status as well as collect and parse BLAST output. In total, 8,626 (44%) gene models had at least one homolog in the NR database, while 7,566 (38.8%) gene models were homologous to at least one protein sequence in the SwissProt database. Detailed homology data for each gene model can be found in **Supplementary Tables 29S** and **30S**.

SM4.1.2 **Gene Ontology Terms**

Gene Ontology (GO) terms (http://www.geneontology.org/) were assigned to gene models with the annot8r annotation pipeline available from the Blaxter Lab, Institute of Evolutionary Biology, School of Biological Sciences, The University of Edinburgh (http://www.nematodes.org/bioinformatics/annot8r/). Our gene models were aligned to the curated subset of protein sequences available via the annot8r interface with NCBI's blastp protein alignment tool. For alignments with an e-value of less than 1e-04, the GO terms associated with the annot8r protein sequence were assigned to the gene model involved in the alignment. Results of the annot8r pipeline were stored in an intermediate form inside a PostgreSQL database installed on a highly-reconfigurable supercomputer (8 x Intel Xeon X7550,

512GB RAM, CentOS 5.5) provided by the Advanced Computing and Information Systems Laboratory at the University of Florida (https://www.acis.ufl.edu/). These results were then transformed into a format suitable for our comparative neurogenomics database, NeuroBase[19], via a custom Perl script. We also developed a Java tool for determining the number of sequences assigned any given GO term, the total number of annotations involving any given GO term, as well as the digital expression level (RNA-seq) of the sequences annotated with any given GO term. The tool performs these analyses by first constructing a tree of the parent-child relationships of the annotated GO terms, utilizing the term2term and term databases provided on the Gene Ontology website. This tool then performs a depth-first traversal of the tree, calculating each of the above statistics at the leaf nodes and propagating the results up the tree. These methods resulted in a total of 61, 265 GO terms being associated to the *P. bachei* gene models, across a total of 9,914 terms.

SM4.1.3 KEGG Pathways

P. bachei gene models were placed in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways by utilizing the annotation capabilities of the zero-click informatics pipeline as described[19,20]. Briefly, homology between the *P. bachei* gene models and the UniProtKB/*Swiss-Prot* (SP) protein database was determined using the methodology described in S.5.1.1. The accession number of each protein homologous to a given gene model is then looked up in Redis keystore, version 2.4.17, containing mappings of SP accessions to KEGG Orthology identifiers and the pathways containing those identifiers. The lookup of each accession produces a list of KEGG annotations for the SP protein, which is then mapped to a gene model based on homology between that gene and the SP protein. Using the above mapping heuristic, a total of 4,414 KEGG annotations were assigned to 3,682 gene models.

SM4.1.4 UniProtKB/Swiss-Prot to KEGG Associations

Associations of SP accessions to KEGG Orthology identifiers (KO) and KEGG pathways were obtained by applying a custom Python module to the genes_ko.list, genes_uniprot.list, ko_pathway.list, and map_title.tab, available from the KEGG FTP server with a paid subscription (http://www.genome.jp/kegg/). The module utilizes genes_uniprot.list to create an association between each SP accession and a gene accession. The module then looks up the gene accession in the genes_ko.list to find the list KO identifiers associated with the gene accession. For each KO identifier, the module constructs a list of pathway identifiers in which that KO identifier (and thus SP protein) is a component, using the ko_pathway.list. Finally, the module looks up the name of each pathway in map_title.tab, producing a tab-delimited file in the format: SwissProt Accession | KO | Pathway ID | Pathway Name.

The module then parses the tab-delimited file and loads each SP accession to KEGG data mapping into a redis keystore (http://redis.io/) for efficient lookup, using SwissProt Accession-KO-Pathway ID as the fields to construct the key. This heuristic results in each SwissProt accession having an entry in the keystore for each pathway in which it can be placed as a component. As a single KO identifier can be mapped to multiple pathways, it can be expected that a protein from SP that is mapped to a single KO term may have multiple entries in the keystore.

SM4.1.5 **Pfam Domains**

Pfam Domains were assigned to *Pleurobrachia bachei* gene models using the Pfam annotation package available from the Sanger Institute (the pfam_scan.pl script ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/README). As per the installation instructions, HMMER 3.0 was downloaded and compiled on the reconfigurable supercomputer discussed in S6.1.2. The Moose Perl framework (http://search.cpan.org/dist/Moose/) and BioPerl 1.6.901(http://search.cpan.org/~cjfields/BioPerl-1.6.901/) were also installed, and the Sanger Institute's Hidden Markov Models (Pfam-A.hmm, Pfam-B.hmm) were downloaded. pfam_scan.pl was then run on our non-redundant set of gene models. After filtering the results of pfam_scan.pl with a custom Perl script, 6,824 gene models had at least one Pfam domain annotation with an expect value less than 1e-04. A Java tool was developed to sum the number of gene models with annotations to each pfam domain.

SM4.2 **Transcriptome Analysis**

SM4.2.1 **Reference Transcriptome Assembly**

SM4.2.1.1 **454 Transcriptome Assembly**

Utilizing the RNA isolation, library construction, and sequencing methods described in S1.3 and S1.5, raw 454 reads were obtained from the following tissues: tentacles, comb plates, the aboral organ, and embryonic stages. Each of these data sets were individually assembled in a two-stage pipeline consisting of version 3.2.1 of the MIRA assembler (command-line parameters: -SK:pr=70 -AL:mo=30:ms=10:mrs=70 -AS:mrl=30: -CL:qc=no -OUT:sssip=yes) and the CAP3 assembler (command-line parameters: -k 0). MIRA and CAP3 were installed on the reconfigurable supercomputer described in S6.1.2. A custom Python script was developed to handle job submission and intermediary data transformation. Quality, vector, and primer trimming were performed using the native capabilities of the MIRA assembler.

Combined assembly was also performed on raw reads from tentacles, combs, and the aboral organ at the University of Florida's Interdisciplinary Center for Biotechnology Research (ICBR). An initial assembly of 454 sequences from these tissues was performed with Newbler assembler version 2.3 (454 Life Science, Branford, CT), after masking and trimming the reads to remove sequencing repeats, primers and/or adaptors used in cDNA library preparation and normalization. A hybrid assembly was then performed by submitting the contigs and singlets produced by Newbler as input to PTA (Paracel Transcript Assembler) version 3.0.0 (Paracel Inc, Pasadena, CA).

In PTA, all sequences were checked and masked for universal and species-specific vector sequences, adaptors, and PCR primers used in cDNA libraries. *Escherichia coli* contamination and mitochondrial and ribosomal RNA genes of *Pleurobrachia* were identified and removed from input sequences using default settings to ascertain the novelty of the sequences. Poly (A/T) tails and intrinsic repeats that were identified in *Homo sapiens*, such as simple sequence repeats and SINE elements, were annotated prior to clustering and assembly. Low base-call quality data

were trimmed from ends of individual sequences and sequences with length <75 bp were excluded from consideration during initial pair-wise comparison. After cleanup, sequences were passed to the PTA clustering module for pair-wise comparison and then to the CAP3-based PTA assembly module for assembly. To efficiently and accurately assemble those input sequences, two assemblies with PTA were conducted on different datasets. The first PTA assembly took input sequences of contigs generated from the Newbler assembly. The second PTA assembly was performed based on the leftover singletons, with all contigs resulting from the first PTA assembly as the reference (or seed) sequences. Summaries of the 454 transcriptome assemblies are presented in **Supplementary Table 3S**.

SM4.2.1.2 **Illumina Transcriptome Assembly**

Utilizing the RNA isolation, library construction, and sequencing methods described for the Illumina data in S1.3 and S1.5, raw reads were obtained from whole adult animals and a population of embryonic animals. Data include 23,240,687,932bp in 236,203,645 sequence reads from adult animals and 22,259,010,820 bp in 228,616,963 sequence reads from *Pleurobrachia bachei* embryonic stages. These two data sets were individually assembled with the SOAPdenovo assembler version 1.05 (command-line parameters: -k 75)[21] on the ACIS supercomputer described in S6.1.2. Initial assemblies contained 54,263,467bp in 202,636 contigs (adult animals) and 17,829,586bp in 69,101contigs (embryonic stages).

Illumina HiSeq sequencing was also performed with RNA extracted from the following ctenophore species: *Euplokamis dunlapae, Coeloplana astericola, Vallicula multiformis, Pleurobrachia pileus, Pleurobrachia sp (*collected from the Middle Atlantic and later identified as a subspecies of *P. pileus* ssp.), *Dryodora glandiformis, Beroe abyssicola, Mnemiopsis leidyi, Bolinopsis infundibulum,* and an undescribed species which belongs to Mertensiidae family. The amount of sequence reads generated for each species can be found in **Supplementary Table 3S**. Each sequencing project was individually assembled using the Trinity *de novo* assembly package [21] on the ACIS supercomputer, using default settings. Prior to assembly, reads for each species were quality trimmed (-m 50 –q 20) and had adapter contamination removed (one mismatch allowed and minimum match length 5) with cutadapt[22]. Full summaries of the Illumina transcriptome assemblies are presented in **Supplementary Table 10S**.

SM4.2.1.3 **Ion Torrent Assembly**

Utilizing the RNA isolation, library construction, and sequencing methods described for Illumina in S1.3 and S1.5, raw Ion Torrent reads were obtained for 1-cell, 2-cell, 4-cell, 8-cell, 16-cell, 32-cell, 64-cell, gastrula, one day, and three day *Pleurorbachia bachei* embryonic stages. A summary for Ion Torrent transcriptome sequencing can be found in **Supplementary Table 3S**. The 1-cell, 2-cell, 4-cell, 8-cell, 16-cell, 32-cell, and 64-cell reads were imported into the CLCBio Genomics Workbench version 5.5 for assembly, using the "Import" option tuned to Ion Torrent reads. Imported reads were then filtered for quality with a cutoff of .05, and an ambiguous base limit of 2 per sequence, given that ambiguous "N" bases are rare in our Ion Torrent sequencing projects. Ion Torrent adapter sequences were removed with the Genomics Workbench's built-in functionality, prior to being assembled with the *de novo* assembly tool

included in the software's toolbox. For each assembly, a minimum length cutoff of 100bp was used with default assembly parameters. To obtain singletons (unassembled reads), original reads were mapped back to the resulting transcriptome assembly using CLCBio's proprietary short read alignment package. Any read failing to align to an assembled transcript was labeled as a singlet and combined with the assembly result to yield the finished assembly.

Gastrula, 1-day, and 3-day transcriptomes were assembled with an alternate method, utilizing the MIRA assembly software, version 3.9.9. First, reads were trimmed for quality with cutadapt, with parameters–m 50 and –q 20. Trimmed reads from each sample were then independently assembled with MIRA, command line options -GE:not=20 IONTOR_SETTINGS -AS:mrl=30:mrpc=1 -OUT:sssip=yes. After running MIRA, the final assembly was produced by concatenating files of contigs, singletons, and discarded reads produced by MIRA during the assembly process. Statistics for each of the Ion Torrent transcriptome assemblies can be found in **Supplementary Table 4S**.

SM4.2.2 **Reference Transcrptome Annotation**

Transcriptomes utilized in this study were annotated with the Moroz lab's zero-click informatics pipeline, Auto-nomics[19,20]. Briefly, the pipeline performed the following steps for annotation. Reference transcriptomes constructed as described in S4.2 depending on sequencing platform, were uploaded by the system to the University of Florida's High Performance Computing Cluster (UFHPC). Each transcriptome was aligned to both NCBI's non-redundant protein database (NR) and the UniProtKB/*Swiss-Prot* (SP) protein database with the software and settings detailed in sections below. The top five hits from each database were retained as homologs to transcripts contained in each reference transcriptome. Gene Ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms were associated with each transcript, utilizing the methods outlined in S4.1.1 and S4.1.5. Pfam domains were assigned to reference transcriptomes by first translating transcripts in all six reading frames. Translated sequences were then split into multiple input files, uploaded to UFHPC, and assigned Pfam domains with pfam_scan.pl. Each reference transcriptome, together with its full set of annotation and expression data, were automatically uploaded by the system to the Moroz lab's comparative neurogenomics database, Neurobase, for downstream analysis and visualization[19,20].

SM4.2.3 **RNA Expression Data (RNA-seq)**

SM4.2.3.1 **Gene Model Expression Levels Across Transcriptomes**

All short read RNA-seq data presented in **Supplementary Table 3S** were quality and adapter trimmed with cutadapt prior to alignment to the final set of *P. bachei* predicted gene models with tblastn from NCBI's BLAST+ software package, using default settings. A custom Python module was developed, which parsed the BLAST output of each alignment and assigned reads to a gene model if the 90% of the read was involved in the alignment, and the region shared between the read and the gene model was greater than 90% identical. The raw expression value for each gene model in each transcriptome was then assigned by summing the number of reads assigned to each gene model by applying the above cutoffs. A summary of raw expression

counts and expression frequency for the *P. bachei* gene models in each of our sets of transcriptomic data is given in **Supplementary Tables 29S and 30S**.

SM4.2.3.2 **454 Transcriptome (RNA-seq)**

Raw expression counts for transcripts obtained using the 454 assembly strategy described in S4.2.1.1 and **Supplementary Table 3S** were calculated using the read placement statistics produced by the Newbler and MIRA assemblers. Both of these software packages produce text files describing the precise number of reads that were used in the construction of each contig. Thus the raw count values for a given transcript can be calculated by collecting a list of contigs and singlets used to construct that transcript and summing the number of reads that were assigned to each contig in the list (naturally, each singlet contributes a value of 1 to the raw count). A custom Python module was developed to parse the read statistics output of Newbler and MIRA and assign raw count expression values to each assembled transcript using the heuristic detailed above. Expression values for transcripts in each 454 transcriptome can be found on UF online database, http://moroz.hpc.ufl.edu/slimebase2/browse.php.

SM4.2.3.3 **Illumina Transcriptome (RNA-seq)**

The short read Illumina RNA-seq (transcriptome) data presented in **Supplementary Table 3S** were first assembled into reference transcripts using the strategy described in S4.2.1.2. Next, the original Illumina short read sequences were trimmed for quality, filtered of adapter sequence, and subsequently aligned to the newly constructed reference transcriptome with bowtie, through the Moroz lab's zero-click informatics pipeline [19,20]. The resulting SAM alignment file was parsed using a custom Python module integrated into the pipeline, assigning a read to a transcript if it met the requirements of having 90% identity to the transcript, having 90% of its length contained in the alignment, and having the transcript in question as it's best unambiguous match. Raw expression counts for each transcript were then taken as the sum of the number of reads whose alignments to the transcript satisfied the requirements mentioned above. All Illumina transcriptome data is available for direct download from UF online database, http://moroz.hpc.ufl.edu/slimebase2/browse.php.

SM4.2.3.4 **Ion Torrent Transcriptome (RNA-seq)**

Reads from the 1-cell, 2-cell, 4-cell, 8-cell, 16-cell, 32-cell, and 64-cell early embryonic stages described in **Supplementary Table 3S** were assembled using the CLCBio Genomics Workbench. The proprietary assembler included in the genomics workbench produces a tab-delimited file containing the number of reads assigned to each transcript. Each transcript present in this file was assigned a raw expression count equal to that number. Additionally, the assembler was instructed to retain a list of reads that could not be placed into any contig. These reads were included in the assembly as singlets, with an assigned raw count expression value of 1.

Developmental transcriptomes from the gastrula stage as well as 1-day, 3-day embryo RNA-seq data (**Supplementary Table 3S**) were assembled using MIRA alone. Like the CLC

Bio Genomics Workbench), MIRA also produces a file containing the number of reads assigned to each transcript. Each transcript identifier was looked up in this file and the number of reads assigned to the identifier was taken as the raw value for the transcript's expression level. Transcripts present in the MIRA assembly file but absent in the read statistics file were assigned the status of singlet, with a raw expression value of 1.

Expression values for all transcripts described in this section can be found online as described in **Supplementary Tables 29S and 30S**.

SM4.2.3.5 **Identification of *Pleurobrachia bachei* Orthologs**

In general, identification of putative orthologs of genes for *Pleurobrachia* started by using BLAST (BlastP and TBlastN) of conserved query sequences against *P. bachei* databases, including the gene models and all the individual adult and developmental transcriptomes (**Supplementary Table S3**). Query sequences were obtained from publically available databases; National Center for Biotechnology Information (NCBI), the Joint Genome Institute (JGI), the Broad Institute, as well as the *A. queenslandica* genome database (www.metazome.net/amphimedon). Proteins were identified by the presence of specific protein domains found by searching the Pfam (''http://pfam.sanger.ac.uk/search''http://pfam.sanger.ac.uk/search) and SMART (''http://smart.embl-heidelberg.de/''http://smart.embl-heidelberg.de/) databases.

Pfam composition[23], Gene Ontology [24], and KEGG (Kyoto Encyclopedia of Genes and Genomes) [25,26] were used to further validate *P. bachei* orthologs on the project database (see below). The *P. bachei* genome browser also integrates expression data tracks of different transcriptomes from adult tissue types and developmental stages.

SM4.2.3.6. **Expression Profiling**

Expression profiling was performed on all our transcriptional data. We have robust expression data due to our sequencing library construction which used a reduced representation strategy. We generated one read per transcript. To determine the expression level of a particular transcript the number of sequence reads per individual transcript is divided by the total number of sequence reads in the entire project. This produces a frequency of expression or expression. Hierarchical clustering was performed in Spotfire. The algorithm used in Spotfire is a hierarchical agglomerative method (http://spotfire.tibco.com/).

SM4.2.3.7 **Prediction of Secretory Peptides**

In order to achieve an open but tractable initial prediction set in the absence of prior studies, we developed three non-overlapping sets of results using different cutoff criteria (**Supplementary Table 26S**). Specifically, we combined multiple distinct heuristic tools in our secretory peptide prediction pipeline, to filter and rank potential secretory products from both genome and transcriptome data. Because almost all pre-secreted peptides are localized to the endoplasmic reticulum via the presence of an N-terminal signal peptide consisting of 5-20 hydrophobic residues, this heuristic forms the strongest marker of secreted products. For signal

peptide prediction, we employed the SignalP[27] and TargetP[28] tools. SignalP utilizes separate Hidden Markov Model (HMM) and Neural Network (NN) components, trained on a variety of organisms, for the sole purpose of secretory signal peptide prediction. TargetP utilizes the Neural Network component of SignalP and separate Neural Networks to discriminate between 'secreted', 'mitochondrial', and 'other' localization signals. To eliminate 'contamination' with putative transmembrane proteins, TMHMM[29] is employed to predict transmembrane topology. Transmembrane domains, like signal peptides, are characterized by a region of hydrophobic residues of similar length, often causing false-positive signal peptide or transmembrane results from trained prediction methods such as TMHMM. Phobius[30] uses complementary HMM, trained separately for transmembrane domains and signal peptides, to call a consensus.

Difference in final predictions between simply removing TargetP criteria, and setting the certainty cutoff to 2 was minimal, due to overlap with SignalP. Additionally, initial tests revealed that changing the target certainly cutoff from 1 to 2 nearly doubled the positive prediction count uniformly across other variables. Thus, it was kept at 1 for all selection sets. Phobis compares the probability scores of transmembrane predictions and signal peptide predictions directly, and so does not produce its own probability score. Thus, adherence to 'signal peptide with no transmembrane domain' is strictly enforced for all predictions. We observed signal cutoff to have the greatest effect on prediction numbers. The default signal and target certainly cutoff 50% for both NN and HMM proved too lenient for our purposes. We examined result density over cutoff percentages to estimate meaningful cutoff values. In order to achieve an open but tractable initial prediction set, we developed three non-overlapping sets of results using the above reasoning (**Supplementary Table 28S** and **32S**).

SM5 Data Availability and Visualization

SM5.1.1 Genome Project Homepage

All data and information on the assembly of the *P. bachei* genome is available online via the *P. bachei* Genome Project homepage, located at http://moroz.hpc.ufl.edu/genomes/genome_pages/pleurobrachia.html . The homepage contains links to the custom UCSC genome browser mirror, comparative transcriptomics database, *Neurobase*, and a download section (all described below), that contains additional direct download links for the data used and presented in this study.

SM5.1.2 UCSC Genome Browser Mirror

A custom mirror of the UCSC Genome Browser was installed and available for public use at homepage http://moroz.hpc.ufl.edu/genomes/genome_pages/pleurobrachia.html). This mirror has been pre-loaded with the current version 1.1 of the *Pleurobrachia bachei* draft genome. Annotation tracks were generated using the blat alignment tool and loaded into the browser using the UCSC Browser's native support for psl files. The reference genome and all data used to generate annotation tracks can be found in the direct download section of the *Pleurobrachiabachei* Genome Project site.

SM5.1.3 **Comparative Transcriptomics Database**

A comparative transcriptomics database, *Neurobase*, has been developed and implemented here to host and present the transcriptome data associated with this and other studies[19,20]. All reference transcriptomes used in this study are publicly available with their associated annotations on *Neurobase* at http://moroz.hpc.ufl.edu/slimebase2/browse.php . Briefly, *Neurobase* organizes transcriptomes into projects and, for each transcript in a project, displays the transcript's homology information, detected expression level, associated GO terms, KEGG pathway placement, and detected Pfam domains. *Neurobase* provides an in-house BLAST server for quick homology search. Additionally, annotations can be searched via keyword at the web interface. Custom GO-tree, KEGG-pathway, and Pfam domain visualization systems have been developed to facilitate the exploration and mining of these annotation families. Finally, *Neurobase* supports exporting each project to a tab-delimited file, containing a full list of that project's transcripts linked to their homology results, GO terms, KEGG pathways, Pfam domains, and expression estimates. The results of BLAST and annotation keyword searches can also be exported to a tab-delimited file for downstream analyses.

SM5.1.4 **Direct Download**

The *P. bachei* draft genome assembly, all reference transcriptomes, current gene predictions, and annotation data sets are available for download at homepage http://moroz.hpc.ufl.edu/

SM6 **Primary Accessions**

Whole-Genome Shotgun projects and trasncriptomes have been deposited at GenBank SRA, genomic data under SRP001155 and transcriptome data under SRP000992.

Annotated RNA-seq assemblies for both *Pleurobrachia* and related ctenophore species are also available at the University of Florida Genome/Transcriptome server project http://moroz.hpc.ufl.edu/ webpages

The original RNA-seq data can be accessed via GeneBank short read archive. Deposited RNA sequences are at NCBI with accession numbers: ELAV-like 1, JN202321; ELAV-like 3, JN202319; ELAV-like 4, JN202320; Sialin-1, JN202322;Sialin-2,JN202323; Sialin-3, JN202324; Dicer JN202325; Argonaute, JN202326; iGluR1, JN202327; iGluR2, JN202328; iGluR3, JN202329 iGluR4, JN202330; iGluR5, JN202331; iGluR6, JN202332; iGluR7, JN202333; iGluR8, JN202334; iGluR9, JN202335; iGluR10, JN202336; iGluR11, JN202337; iGluR12, JN202338, iGluR13, JN202339; Ago2, JN798505; Ago3, JN798506; Ago4, JN798507; Piwi1,JN798508; Piwi2, JN798509; Piwi3, JN798510;HEN1 (small RNA 2'-O-methyltransferase), JN798511;Neurexin, JN798512, DNMT, 1575986; Secretory peptides: Mucin-like, JQ700309; GVEDin, JQ700310; SP 3, JQ700311; SP 4 , JQ700312; SP 5, JQ700313; SP 6, JQ700314; SP 7, JQ700315; SP 8, JQ700316;SP 9, JQ700317; SP 10, JQ700318; SP 11, JQ700319; SP 12, JQ700320; SP 13, JQ700321; SP 14, JQ700322; SP 15, JQ700323; SP 16, JQ700324; SP 17, JQ700325; SP 18, JQ700326; SP 19, JQ700327; SP 20,

JQ700328; SP 21, JQ700329; SP 22, JQ700330; SP 23, JQ700331; SP 24, JQ700332; SP 25, JQ700333; SP 26, JQ700334; SP 27, JQ700335; SP 28, JQ700336; SP 29, JQ700337; SP 30, JQ700338; SP 31, JQ700339; SP 32, JQ700340; SP 33, JQ700341; SP 34, JQ700342; SP 35, JQ700343; SP 36, JQ700344; SP 37, JQ700345; SP 38, JQ700346; SP 39, JQ700347; SP 40, JQ700348; SP 41, JQ700349; SP 42, JQ700350; SP 43, JQ700351; SP 44, JQ700352; SP 45, JQ700353; SP 46, JQ700354; SP 47, JQ700355; SP 48, JQ700356; SP 49, JQ700357; SP 50, JQ700358; SP 51, JQ700359; SP 52, JQ700360; SP 53, JQ700361; SP 54, JQ700362; SP 55, JQ700363; SP 56, JQ700364; SP 57, JQ700365; SP 58, JQ700366; SP 59, JQ700367; SP 60, JQ700368; SP 61, JQ700369; SP 62, JQ700370; SP 63, JQ700371; SP 64, JQ700372; SP 65, JQ700373; SP 66, JQ700374; SP 67, JQ700375; SP 68, JQ700376; SP 69,JQ700377; SP 70, JQ700378; SP 71, JQ700379; SP 72, JQ700380

SM7 Phylogenomic Analyses

SM7.1 Methods for Phylogenomic Analyses

To reconstruct basal metazoan phylogeny, we conducted two sets of phylogenomic analyses. All analyses included new data from *Pleurobrachia bachei* and the sponges *Sycon* (Calcarea) and *Aphrocallistes* (Hexactinellida) as well as existing data **Supplementary Table 12S**.

For the first set of analyses, Ctenophora was represented by two species of *Pleurobrachia* and *Mnemiopsis*. For a subsequent analysis, sampling within Ctenophora was expanded to include ten additional taxa. Methods for expanded sampling are described in S7.2. Other minor changes were made to improve the taxon sampling. Additionally, we modified our bioinformatic pipeline to employ a more accurate translation utility, include stricter screening of potential paralogs, and require more taxa to be sampled per gene.

Gene models (predicted transcripts) for novel data herein assembled as described above and publicly available data were handled as in Kocot et al. (2011)[31]. "Unigenes" (contigs from Illumina assemblies and contigs plus high-quality singletons from 454 and Sanger assemblies) from each transcriptome and predicted transcripts from each genome were translated using ESTScan[32]. Orthology determination employed HaMStR local 7[33], which search translated unigenes against the 1,032 single-copy orthology groups (OGs) of HaMStR's "model organism" profile hidden Markov models (pHMMs) set. Unigenes resulting in positive matches where reciprocally compared the *Drosophila* proteome using BLASTP. If the best match was the *Drosophila* protein contributing to the pHMM, the unigene was placed in that OG. To reduce the amount of missing data per taxon, chimerical Operational Taxonomic Units (OTUs) were created for *Acropora*, *Aiptasia*, *Oscarella*, and *Sycon*. Subsequently, OGs with sequences from fewer than fifteen OTUs were discarded. Because ends of sequences can be mistranslated due to low coverage, sequences were trimmed if an X (a codon with an ambiguity, gap, or missing data) occurred within 20 positions of the sequence end. All characters from the X to the end were as missing data. Sequences shorter than 100 amino acids (AAs) after this trimming step were deleted. All remaining sequences in each OG were then aligned with MAFFT[34] using default settings.

In cases where an OTU was represented by more than one sequence in an OG, the sequence selection option of SCaFoS 1.30[35] was employed (puz=yes, gamma=yes, t=5, o=gscl, m=25) to avoid inclusion of paralogs. Sequences were considered "complete" if at least 75% of the sequence was non-missing data characters (characters other than -, X, or ?). If one complete sequence was sampled for an OTU, any sequences with <75% completeness were discarded. If an OTU had two or more complete sequences which where > 5% divergent, all sequences from that OTU were excluded. If two or more acceptable sequences were sampled from the same OTU and were less than 5% divergent, pairwise distances among all sequences in the OG were calculated in TREE-PUZZLE using a gamma parameter with four rate categories and the complete sequence from the OTU with the shortest average pairwise distance to all other sequences in the alignment was retained. Lastly, if only incomplete sequences were sampled, they were combined into one, more complete "chimerical" sequence.

To remove uninformative and ambiguously aligned positions in the resulting alignments, trimming was performed with Gblocks[36] (b1 = b2 = half of the number of sequences in the alignment plus one, b3 = 8, b4 = 10, b5 = h, t = p, p = n) and a custom script to delete any positions with two or fewer non-gap characters. Subsequently, in order to remove highly divergent sequences likely to be paralogs or replete with multiple substitutions, a consensus sequence was inferred for each alignment. For each sequence in each single-gene amino acid alignment, the percent of positions of that sequence that differed from the consensus of the alignment were calculated using the EMBOSS program infoalign's "change" calculation. Any sequence with a 'divergence' "change" value greater than 75 was discarded. Uninformative positions were further trimmed using Aliscore and Alicut[37]. Resulting alignments shorter than 100 AAs were discarded and sequences shorter than 100 AAs were deleted. Lastly, in order to reduce the overall amount of missing data and keep the dataset at a manageable size, OGs with fewer than fifteen taxa and OGs lacking a sequence for *Pleurobrachia bachei* were discarded.

This resulted in a final matrix of 170,871 amino acid positions across 586 genes with 44 taxa. On average, taxa in this matrix are sampled for 51.3% of genes and overall matrix completeness is 45.4% comparable to other phylogenomic studies (e.g., Dunn et al. 2008 [38]). The matrix was analyzed on the Auburn University Molette Lab SkyNet supercomputer using maximum likelihood in RAxML 7.2.7[39] under the WAG+CAT+F model which has been demonstrated to perform similarly the best-fitting amino acid substitution model without the computational cost[31]. Nodal support was assessed using 100 bootstrap replicates. Fungi were used to root the resulting tree (**Extended Data Figure 3a**). Leaf stabilities and branch attachment frequencies were examined using Phyutility[40].

Leaf stabilities were calculated using Phyutility[88]. Because leaf stability scores for the placozoan *Trichoplax* and the nematode *Caenorhabditis* were very low (0.8485 and 0.7998, respectively) and branch attachment frequencies showed that these taxa were not stable among bootstrap replicates (data now shown), the data matrix was reanalyzed (same methods) excluding these two taxa (**Extended Data Figure 3b**). Notably, leaf stability scores for hexactinellid and calcareous sponge OTUs were also low but branch attachment frequencies showed that placement varied primarily within Porifera (data not shown) and thus they were not excluded. .

Further, distant outgroups (Fungi, *Amoebidium*, *Sphaeroforma*, and *Capsaspora*) were excluded to examine the possibility of homoplasy caused by long-branched outgroups. This also

resulted in Ctenophora being placed as a sister group to all other Metazoa with 100% bootstrap support (**Extended Data Fig. 3c**).

Alternative hypotheses of basal metazoan phylogeny were evaluated using the SH-test[89] as implemented in RAxML with the PROTGAMMAWAGF model.

Additional details of the large-scale phylogenomic analyses[88] using different taxon sampling are presented in the **Supplementary Data section SD4**.

SM7.2 Analyses with Expanded Ctenophora Sampling

Taxon sampling for the expanded phylogenomic analysis is presented in **Supplementary Table 13S.** Briefly, we built on the analysis above that used only choanoflagellate outgroups by adding relatively deeply sequenced Illumina transcriptomes for additional ctenophores. A small number of publicly available datasets were also added, in some cases replacing smaller datasets sampled in our earlier analysis.

Methods use to generate and analyse this data set were conducted as described above except where noted in this section. For ctenophores not used in the analyses described above, raw paired-end (PE) Illumina reads were digitally normalized using the script khmer[41] (normalize-by-median.py -C 30 -k 20 -N 4 -x 2.5e9). Singleton reads were deleted and read pairing restored with a pipeline that included scripts provided with khmer. Remaining PE reads were assembled on the Auburn University Molette Lab SkyNet server using the October 2012 version of Trinity[21] using the default settings. For Illumina data, only contigs were processed with TransDecoder for translation (https://sourceforge.net/p/transdecoder/). For Sanger and 454 data, both contigs and high-quality singletons (i.e., unigenes) were translated with TransDecoder. Orthology determination was performed as in S7.1 except OGs sampled for fewer than twenty taxa were discarded which aided subsequent paralog determination (see below). Sequences were then subject to trimming for misaligned ends, removal of short sequences, and MAFFT alignment as described above. Next Gblocks[36], removal of gap positions, and trimming of short sequences were carried out as in S7.1. OGs with fewer than ten taxa were discarded.

To screen groups for evidence of paralogy, an approximately maximum-likelihood tree was inferred for each remaining alignment using FastTree 2[5] using the "slow" and "gamma" options. A phylogenetic tree-based approach was then employed to screen each single-gene tree for evidence of paralogy. First, nodes with ≤0.50 support were collapsed into polytomies. Next, using the custom script PhyloTreePruner (Kocot and Citarella et al., accepted manuscript), the maximally inclusive subtree was retained where all taxa were represented by zero or one sequence(s) or, in cases where more than one sequence was present for one or more taxa, all sequences from a given taxon were monophyletic or part of the same polytomy. Polytomies containing more than one sequence from two or more taxa were permitted. Putative paralogs (sequences falling outside of this maximally inclusive subtree) were then deleted from the input alignment. In cases where multiple sequences from the same taxon formed a clade or were part of the same polytomy, all sequences but the longest (presumably the most complete splice

variant) were deleted. Lastly, in order to reduce the amount of missing data in the final data matrix, any groups sampled for fewer than 20 taxa were discarded.

The remaining OGs were then concatenated using FASconCAT (http://zfmk.de/web/Forschung/Abteilungen/AG_Wgele/Software/FASconCAT/index.en.html). This resulted in a final matrix of 22,772 amino acid positions across 114 genes with 60 taxa. Overall matrix completeness was 51.8%. This matrix was analyzed using ML in RAxML 7.2.7[39] as described in S7.1. Topological robustness for the ML analysis was assessed with 100 replicates of nonparametric bootstrapping, treating any BP $\geqslant$70 as significant nodal support[10].

Alternative hypotheses of basal metazoan phylogeny were evaluated using the SH-test[89] as implemented in RAxML with the PROTGAMMAWAGF model.

SM7.3 Analysis of Individual Genes and Families

Alignments of all orthologs were performed with either ClustalX2 [42-44] or Muscle [45] then, if appropriate, either trimmed manually or trimmed using GBlocks [46] for further comparative analysis of individual gene families. Pfam composition[23], Gene Ontology[24], and KEGG (Kyoto Encyclopedia of Genes and Genomes) [25,26] were used to further validate *P. bachei* orthologs. Once alignments were obtained, basic gene trees were constructed in MEGA 5 [47]. The molecular phylogenetic analysis used the Maximum Likelihood (ML) method based on the Whelan and Goldman (WAG) matrix-based model. The bootstrap consensus tree was inferred from 100 replicates. All positions containing gaps and missing data were eliminated. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of amino acid substitutions per site. Larger phylogenetic analyses were conducted using ML in RAxML 7.2.7 [39]. For ML analyses, the CAT +WAG + F model was used. Topological robustness (i.e., nodal support) for all ML analyses was assessed with 100 replicates of nonparametric bootstrapping.

SM7.4 Analysis of Gene Gain and Gene Loss

Analyses of gene gain and gene loss have been performed using custom scripts as described elsewhere[48]. Briefly, for our study, in addition to the genome of *Pleurobrachia bachei*, we have selected sequenced genomes of several representative animal phyla from basal Metazoans such as *Amphimedon* (Porifera), *Trichoplax* (Placozoa), *Nematostella* (Cnidaria), and representative of key bilaterian lineages: two protostomes such as *Daphnia* (Arthropoda) and *Lottia* (Mollusca); and four deuterostomes such as *Strongylocentrotus* (Echinodermata), *Saccoglossus* (Hemochordata), *Homo* and *Branchiostoma* (Chordata). We have also selected single-cell eukaryotic genomes of *Monosiga* and *Capsaspora* representing potential sister animal taxa and a plant genome of *Arabidopsis thaliana* as an out-group. The sources and sizes of data are outlined in the **Supplementary Table 14S**. In some of these genomes gene models predicted *ab initio* are often short, fragmented and contaminated by translated non-protein coding fragments. To make data more consistent we used all unfiltered gene models for all genomes, even in cases of finished genomes for which refined sets of protein-coding genes are available. Gene gain and gene loss in a group of distantly related genomes has been estimated by the workflow outlined in Ptitsyn and Moroz, 2012[48]. Initial input of our analysis pipeline consists of predicted protein models for a group of 12 selected genomes. Resulting non-redundant sets of

gene models have been compared to each other using reciprocal BLAST with the same similarity threshold (estimated by e-value equal to or lower than 0.0001).

SM7.5 Phylogenomic Analysis Taking Into Account Divergence of Ctenophore Proteins

Both protein alignments previously used to study eukaryotic relationships[49,50] and unpublished phylogenetic markers were added to the 548 alignments previously mentioned, leading to an initial dataset of about 650 phylogenetic markers. All of these alignments were screened to keep alignments that are conserved in animals and their unicellular relatives and those having a simple gene history (i.e. simple orthologous relationships). At this step, 350 alignments were retained. Then, alignments were aligned using Muscle[51] and then trimmed by Gblocks[46] using default parameters except for the maximum proportion of gaps per position (increased from 0% to 20%). Finally, alignments were inspected to correct misaligned portions of sequences. In order to select the most conserved phylogenetic markers, we estimated the saturation of each individual alignment by comparing p-distances (inferred by RAxML[39] under the PROTGAMMALG model) and uncorrected pairwise distances, and calculating the slope of the regression line. The 225 alignments having the highest slope values (i.e. the most conserved alignments) were kept for further analyses (all data are shown in the Excel Table 31S).

To detect outliers, a search for the best ML tree and the analysis of 100 bootstrap replicate were performed from each trimmed alignment using RAxML version 7.2[39] under the PROTGAMMALG model. These trees were then combined to detect outliers (i.e. contaminants, lateral gene transfers or hidden duplications) by Phylo-MCOA[52] using nodal distances and a threshold value equal to 70. Finally, all detected outliers were manually removed from their corresponding alignments.

To estimate divergence of ctenophore proteins all 225 alignments were ranked according to the evolutionary rate of ctenophore proteins. For this purpose, anthozoan, desmosponge and ctenophore proteins were extracted from each alignment, and patristic distances between these species were inferred by RAxML under the PROTGAMMALG model using a fixed topology with different phylogenetic hypotheses. We then calculated the ratio 'R' of the average patristic distances between ctenophores and the last common ancestor of animal, to the average patristic distances between anthozoans/demosponges and the last common ancestor of Metazoa. This metric, R, was used to assess the divergence of ctenophore proteins and rank the 225 alignments (R values and the rank for each alignment are given in the **Supplementary Table 33S**. Finally, alignments were concatenated into several supermatrices using a custom-made script.

Bayesian inferences analyses were performed with the CAT-Poison + Γ4 mixture model using the -dc option implemented in the program PhyloBayes version 3.2[53]. For each dataset, one run was performed with a total length of 15,000 cycles. The first 6,000 points were discarded as burn-in, and the posterior consensus was computed by selecting one tree every ten. Maximum Likelihood analyses were performed by analyzing under the PROTGAMMALG, for each dataset, 100 bootstraps replicates with the rapid BS algorithm model implemented in RAxML. Node supports for the alternative positions of ctenophores were calculated using the ETE package[54].

SM8 Analysis of DNA Methylation in *Pleurobrachia*

We used ELIZA based colorimetric assays and quantified global 5-methylcytosine (5-mC) and hydroxylmethyl-cytosine (5-hmC) in the *P. bachei* genome[57]. We isolated genomic DNA (see section S1.2) from 6 individual adult *P. bachei* and 3 rat brain/hippocampal tissue (positive control). Quality and quantity of genomic DNA was analyzed on a Qubit® 2.0 Fluorometer (Life technologies) and an agarose gel. The same quantified genomic DNA samples were used for both 5-mC and 5-hmC assays. We used the MethylFlash Methylated/Hydroxymethylated DNA Quantification Kit from Epigentek to quantified global concentrations of 5-mC and 5-hmC in the genomic DNA samples (Cat # P-1034 and Cat # P-1036, Epigentek, Farmingdale, NY). Here, the gDNA is immobilized to wells coated with a specific DNA affinity substratum and the bound methylated fraction of the gDNA is then recognized by a specific antibody for 5-mC or 5-hmC and quantified through an ELISA-like reaction. Absorbance was measured at 450 nm. A diluted positive control is used to generate a standard curve from which the 5-mC or 5-hmC concentration is obtained. Slope of the standard curve is determined using linear regression and the percentage of 5-mC or 5-hmC in the total DNA is calculated using the following formula and absolute quantification of 5-mC and 2hmC were determined as

$$\text{5-mC\% or 5-hmC\%} = \frac{\text{Sample OD} - \text{Negative Control OD}}{\text{Slope} \times \text{Input DNA Amount} \times \text{N*}} \times 100\%$$

N* factor to normalize 5-mC (2) or 5-hmC (5) in the positive control to 100%.

Three replicates were run for every sample then the whole assay repeated using three different times. Absolute quantification of 5-mC and 5hmC were determined and data are reported as an mean ± S.E.M. Data was then plotted in Graphpad Prism 6 (http://www.graphpad.com/).

SM9 Molecular Cloning

SM9.1 cDNA Library Construction

High quality RNA was obtained as described (see RNA isolation, Section S1.3), and used as starting materials both for RNA-seq expression analysis and generation of complementary DNA (cDNA) libraries. We constructed 5' and 3' tagged libraries that were amplified and then used for cloning genes of interest or for Rapid Amplification of cDNA Ends (RACE). cDNA libraries were constructed from major adult tissues (aboral organ, combs, tentacles, mouth, stomach/pharynx) as well as embryonic and early developmental stages (from 1-cell, 2-cells, 4-cells, 8-cells, 16-cells, 32-cells, 64-cells, early and later gastrula, 1 to 3-4 days embryos and larvae) of *P. bachei*. We generate cDNA using SuperScript® II Reverse Transcriptase (Cat #18064-014, Life Technologies) or Smartscribe Reverse Transcriptase (Cat # 639537, Clontech). From this point in construction, we followed one of two methods. In the first method, we added a template switching oligo as in the SMART™ cDNA Library Construction Kit (Cat# 634901, Clontech) to generate 2nd strand DNA and then amplified the library. Alternatively, we polished the generated cDNA with T4 DNA polymerase then ligated a 5' adaptor before amplifying the library as described in the Marathon kit (Cat # 634913, Clontech).

SM9.2 **Cloning**

We generated specific primers (Integrated DNA Technologies (IDT), Iowa) for genes of interests and PCR was used to amplify selected fragments. The templates for the PCR amplification were cDNA libraries from the various tissue types in the adult *P. bachei* or developmental stages. Manufacture's protocols were followed for various Taq polymerases (TaKaRa LA Taq, Cat # RR002M, Clontech or Platinum® PCR SuperMix High Fidelity, Cat # 12532-016, Life technologies). PCR products were cloned into pCR™4-TOPO® TA vector and transformed into TOP 10 cells (Cat # K4575-01, Life technologies). *E.coli* colonies were grown up and plasmid DNA were extracted (Cat# 19064, Qiagen), then sequenced using either local University of Florida facilities or SeqWright, Houston, TX.

SM10 *In situ* Hybridization

Once a gene of interest had been cloned, *in situ* hybridization was performed to localize the message RNA of this gene in *P. bachei*. We modified the non-radioactive chromogenic *in situ* hybridization protocol from several published protocols[58-60]. The employed method is based on immunological detection using alkaline phosphatase-conjugated antibodies. We used digoxygenin (DIG)-labeled antisense RNA probes with nitro blue tetrazolium chloride/5-bromo-4-chloro-3-indolyl phosphate (NBT/BCIP) as the alkaline phosphatase substrate for mRNA detection. DIG-labeled uridine triphosphate (UTP) incorporates into the antisense RNA probe when it is synthesized. When the probe was hybridized to whole-mount ctenophore preparations, the preparations are incubated with alkaline phosphatase-conjugated anti-DIG antibodies. Dephosphorylation of alkaline phosphatase substrate causes a colored precipitate to form; in the case of NBT/BCIP it is blue. All the preparations were done in an RNAse-free environment. The detailed description of the *in situ* hybridization methods were previously reported[61]

SM11 Immunohistochemistry

Animals were fixed overnight in 4% paraformaldehyde in 0.1 M phosphate-buffered saline (PBS) at +4º C, washed for 2 hours in 0.2 M phosphate buffer saline (PBS, pH=7.6) and dissected as necessary for analysis. The tissue was dehydrated for 5 minutes in 100% methanol at -10º C, rinsed in PBS containing 0.01% Triton X-100 (PBT) for 2 hours and then pre-incubated overnight in a blocking solution of 6% goat serum in PBT. Then the tissue was incubated for 48 hours at +4º C in the primary antibodies diluted in 6% goat serum at a final dilution 1:40 (Serotec, Cat# MCA77G, Rat Anti Tubulin Alpha). Following a series of PBS washes for 6 hours, the tissue was incubated for 12 hours in secondary antibodies at a final dilution 1:20 (Molecular Probes, Cat# A11077 Alexa Fluor 568 goat anti-rat lgG or Cat# A11006 Alexa Fluor 488 goat anti-rat lgG). To label muscle fibers in ctenophores we used a well-known marker Phalloidin (Molecular Probes, Cat# A12379 Alexa Fluor 488 phalloidin or Cat# A12380 Alexa Fluor 568 phalloidin). Tissue was fixed in 4% paraformaldehyde, washed in PBS and incubated in Phalloidin solution in PBT for 4 to 8 hours at a final dilution 1:80. Then samples were washed in several PBS rinses for 6 hours. In double-labeling experiments the fluorescent signals were differentiated by using fluorescent labels of different wave length: 488 and 568. After washing in PBS, the samples were mounted in Fluorescent Mounting Media (KPL, Cat# 71-00-16) on glass microscope slides to be viewed and photographed using a Nikon Research Microscope Eclipse

E800 with Epi-fluorescence, and BioRad (Radiance 2000) Laser Scanning confocal microscope at Friday Harbor Laboratories (courtesy of Dr. Victoria Foe).

In addition, we used three selected ctenophore species (*Pleurobrachia, Bolinopsis* and *Beroe*) and screened them using commercially available antibodies (Instar & Sigma) against serotonin, histamine, GABA, FMRFamide, and pedal peptide (known to control cilia movement in a wide array of invertebrate species). None of the employed transmitter markers (save FMRFamide) label any neurons in each of these three ctenophore species (both using whole mounts and small microdissected areas from different structures); more than 50 animals (>500 preparations) were used for various modifications of protocols and different fixatives, antisera dilutions, various sources of anitibodies and fluorescent reporters. In contrast, the same antibodies worked perfectly as specific markers of various neuronal populations in representatives of all other phyla we tested (including echinoderms (holothurians, sea urchins), molluscs, annelids, nemertines, Chaetognatha, acoela, and cnidarians). In *Pleurobrachia*, only Sigma FMRFamide antibody did label a limited population of small neuron-like cells (not shown) in the areas of apical polar fields – structures involved in coordination of locomotion. However, no RFamide related genes were identified in the sequenced *Pleurobrachia* genome. At the same time, GABA antisera specifically labeled a subpopulation of muscle cells and mesenchymal cells located in mesoglea (but not in areas where neurons were located using α-tubulin Serotec antisera).

SM12 Scanning Electron Microscopy (SEM)

Animals were fixed in 2.5% glutaraldehyde in 0.2 M phosphate-buffered saline (pH=7.6) for 3-4 hours at room temperature, and washed for 2 hours in 2.5% Sodium Bicarbonate. For secondary fixation, we used 2% osmium tetroxide in 1.25% Sodium Bicarbonate for 2-3 hours at room temperature. Tissue was then rinsed 3 times in distilled water and dehydrated in Ethanol. Tissue samples in 100% Ethanol were then placed for drying in Samdri-790 Critical Point Drying (Tousimis Research Corporation). After the drying process, samples were placed on holding platforms and processed for coating on Sputter Coater (SPI Sputter). SEM observations and recordings were done on NeuScope JCM-5000 microscope (courtesy of Dr. Adam Summers).

SM13 Electrophysiological Methods

SM13.1 Preparation

Before experiments, animals (*Pleurobrachia bachei*) were maintained in sea water at 4-6° C for up to 7 days. In most cases, to get access to isolated muscle cells small sections of ctenophore body around comb plates containing muscle cell layers and mesoglea were mechanically triturated in 400-500 µL artificial sea water (ASW, see Solutions below). Dissociated muscle cells were plated on 35-mm Petri dishes filled with ASW and allowed to settle for at least one hour before recording. Petri dishes with cells plated were kept at 4^{o} C and used for experiments within about 4-5 h. Smooth muscle cells were identified morphologically - elongated shape, a number of processes and, in most cases, functionally - ability to contract either spontaneously or being stimulated with high potassium solution (intracellular solution, see Solutions) applied focally through the recording pipette. In most cases shortened lengths (50-300 µm) of muscle fibers were used in the experiments to provide better space clamp in whole-cell recordings.

SM13.2 Electrophysiological Recording and Analysis

Patch electrodes for extracellular (spikes) and whole-cell recordings were pulled from borosilicate capillary glass (Sutter Instrument, BF150-86-10) using a Flaming-Brown micropipette puller (P-87, Sutter Instruments). All currents were recorded using an Axopatch or 200B amplifier controlled by a Digidata 1322A and pClamp 9.2 and data sampled at 5 kHz. Action potentials (APs, spikes) were recorded in track mode using cell-attached loose-patch configuration. Whole-cell currents were recorded in voltage clamp mode (amplifier configuration – whole cell b=0.1) at a holding potential of -70 mV unless otherwise noted. Bath solution change or neurotransmitter candidate application for both extracellular AP and whole cell recordings were performed with a rapid solution changer, RSC-160 (Bio-Logic - Science Instruments, Claix, France). Neurotransmitter candidates were applied to muscle cells in 100-1000ms duration pulses. In some experiments pulse duration was used to regulate neurotransmitter concentration. Experiments were carried out at ~21°C. Data were analyzed with Clampfit 9.0 (Molecular Devices) in combination with SigmaPlot 10.0 (SPSS Inc., Chicago, IL, USA). The amplitude of the muscle cell response was taken/estimated as the mean number of spikes within 5-10 s after the beginning of the stimulation and normalized to the maximal response. Five-15 responses were averaged for every intensity of each neurotransmitter candidate. All results are expressed as means ± s.e.m. of *n* observations. For videomicroscopy dissociated cells were plated on glass coverslips coated with Concanavalin A (Sigma-Aldrich) and allowed to attach for at least 15 min. Time-lapse series were acquired with QImaging EXi CCD camera using DIC mode of Nikon Eclipse 2000 inverted microscope.

SM14 Calcium Imaging

Petri dishes with ctenophore muscle cells plated were placed on the stage of an inverted microscope (Olympus IX-71) equipped with a cooled CCD camera (ORCA R2, Hamamatsu). Cells were injected with calcium sensitive probe (Fluo-4, ~5mkM) through patch pipette. Fluorescence imaging was performed under the control of Imaging Workbench 6 software (INDEC Systems). Stored time series image stacks were analyzed off-line using Imaging Workbench 6, Clampfit 10.3, SigmaPlot 10/11 or exported as TIFF files into ImageJ 1.42 (available from public domain at http://rsbweb.nih.gov/ij/index.html).

SM15. Solutions and Chemicals

For action potentials (APs) and whole-cell recordings, artificial sea water (ASW) containing (in mM): 460 NaCl, 10-20 KCl, 11 $CaCl_2$, 55 $MgCl_2$, 10 HEPES was used to bath ctenophore muscle cells. In some cases we also used low magnesium sea water saline contained (mM) 486 NaCl, 10-20 KCl, 13.6 $CaCl_2$, 9.8 $MgCl_2$ and 10 HEPES; 0 Mg^{2+} seawater saline (mM, NaCl 486, KCl 5, $CaCl_2$ 13.6, Glucose 30, HEPES 10) and seawater saline without divalent cations: 486 NaCl, 20 KCl, 70 Glucose, 10 HEPES, 1 EGTA, 1 EDTA. ASW also fill the pipette electrode for extracellular action potential recordings. The whole-cell pipette solution contained (in mM): 210 KCl, 0-0.1 $CaCl_2$, 1-2 EGTA, 696 Glucose, 10 HEPES. For calcium imaging experiments intracellular solution contained (mM): KCl 210; GTP 0.5, ATP 0.5, $MgCl_2$ 1, Glucose 696, EGTA 1, HEPES 10. In some experiments 210 mM KCl was replaced with 210 mM CsCl. The pH of solutions was adjusted with NaOH or Trizma base (Sigma) to 7.7-7.9.

Following compounds were tested as possible neurotransmitter candidates (stock solution concentration/final maximal concentration tested, mM): L-Glutamate (Sigma, 500/5); D-Glutamate (Sigma, 1000/5); L-Aspartate (Sigma, 1000/5); D-Aspartate (Sigma, 1000/5); Acetylcholine (Sigma, 500/1); L-Arginine (Sigma, 1000/1); Dopamine (Sigma, 500/1); GABA (Sigma, 500/1); ATP (Sigma, 200/0.5); Taurine (Sigma, 1000/1); L-Glycine (Sigma, 100/0.5); Histamine (Sigma, 200/1); DEA/NO (Sigma,100/0.2); PHE-MET-ARG-PHE amide (Sigma, 50/0.2).

SM16 Pharmacological Assays

Small intact animals were tightly pinned to a Sylgard-coated Petri dish with small steel needles to prevent any body movements other than cilia beating. Large animals were dissected, and part of a body wall with 2-3 cilia rows was pinned the same way to the Petri dish. To record cilia movements we used glass microelectrodes filled with 2M potassium acetate with resistances of 5-20 MΩ. The tip of the electrode was carefully placed next to the cilia combs with the help of micromanipulators and under visual control via dissecting microscope, so that during cilia beating cilia was touching the tip of the electrode. This physical contact created a brief electrical signal that was picked up by A-M System amplifiers (Neuroprobe 1600) and recorded on paper and in digital form using Gould Recorder (WindoGraf 980). Thus each cilia beat was translated into a fast electrical spike, which allowed a digital recording of cilia beat frequency, but not the amplitude and power of beating.

To test the possible role of different known neurotransmitters in cilia control, we bath applied them into the recording dish using a graduated 1ml pipette attached to a long small-diameter tube. Final concentrations were calculated from the known volume of injected solution and the known volume of the recording dish. The following signal molecules had been used in these experiments: GABA (Sigma, 0.01-0.5 mM), acetylcholine (Sigma, 0.01-0.5 mM), serotonin (Sigma, 0.01-0.5 mM), glutamate (Sigma, 0.01-0.5 mM), dopamine (Sigma, 0.01-0.5 mM), histamine (Sigma, 0.01-0.5 mM), glycine (Sigma, 0.01-0.5 mM), aspartate (Sigma, 0.01-0.5 mM), octopamine (Sigma, 0.01-0.5 mM), FMRFamide (Sigma, 0.02-0.2 mM), cAMP (Sigma, 0.02-0.2 mM), cGMP (Sigma, 0.02-0.2 mM), NOC-9 (Calbiochem, 0.05-0.4 mM), diethylamine NONOate (Sigma, 0.05-0.4 mM). To understand whether the possible effect was direct on the cilia cells or indirect via other interneurons, chemical isolation was used by bathing the preparation in high Mg^{2+} saline for 5-15 minutes (333mM magnesium chloride was added to filtered seawater at 1:1 ratio). The total number of preparations used in this study was 120.

Screening for transmitters controlling cilia. We performed pharmacological screening to test whether known classical low molecular weight neurotransmitter act on the ciliated cells in *Pleurobrachia*. There were two control preparations: (i) isolated effectors (combs or microdissected ciliated cells) with tonic cilia beating frequencies (5-10 Hz); and (ii) semi-intact preparations with 2-4 intact comb rows under control of the aboral organ mediating a series of coordinated patterns of bursting activity. In both cases, cilia movements were recorded using a glass microelectrode positioned next to the cilia. The microelectrode were connected to an amplifier and every strike of the cilia row will be recorded as an electrical signal as described above. We have already established that stable patterns of cilia beating can be maintained for

>10-12 hrs under these conditions (n=12). Tested compounds were applied into the recording Petri dish. First, we tested 2 different concentrations of a given compound (1 mM and 0.05 mM, at least 3 preparations each). Then, if any statistically significant effect was recorded after analyzing the frequency and pattern of cilia beating, a more detailed dose-dependent analysis followed to establish a threshold and compare it with results of microchemical measurements to estimate the physiological range of concentrations for a transmitter candidate. Application of fresh filtered seawater always served as the control.

We made preliminary tests on 19 compounds to examine their role in cilia beating control (0.1 and 1.0 mM; 3 times each). These included serotonin (5-HT), acetylcholine (ACh), L-DOPA, dopamine (DA), noradrenaline (NA), adrenaline, octopamine, L-/D- glutamate, L-/D-aspartate, GABA, glycine (Gly), histamine (His), ATP, ADP, AMP, Pedal Peptide (known to accelerate cilia beating in bilaterians), and taurine. This screening indicated that none of these compounds affect cilia beating frequency.

SM17 Determination of the Presence of Classical Neurotransmitters by Capillary Electrophoresis (CE)

SM17. 1 Overview of CE methods

Multiple CE separation techniques were employed to analyze tissue extracts for the presence of a number of neurotransmitters including (see **Supplementary Tables 22S-25S**). While all these methods used CE separations, complimentary detection methods, including laser induced fluorescence (LIF), laser-induced native fluorescence (LINF)[62] and electrospray ionization mass spectrometry (ESI-MS)[63,64], were used to ensure broad coverage and low detection limits for the specific analytes of interest. For example, the CE/LIF system is well suited for a broad range of analytes; the CE-LINF provides excellent figures of merit for natively fluorescent molecules such as the indoleamines and catecholamines, and the CE EIS-MS confirmed the identities of many analytes given the chemically rich mass spectrometry detection.

Tentacles, combs, tentacular sheaths and stomach from *Pleurobrachia bachei* were individually removed using dissection scissors under a 10x stereoscope and immediately rinsed with ultrapure water. Analytes were extracted by adding a volume of approximately 2x the dissected tissue sample of extraction media consisting of 49.5/49.5/1, methanol (LC-MS grade)/water/glacial acetic acid (99%) by volume and homogenized by 1 min of sonication. Samples were allowed to extract at 4°C for 90 min and then centrifuged at 15,000 RPM for 15 min and supernatant removed and frozen at -80°C until analysis.

The CE-LINF instrument employed deep ultraviolet excitation of catecholamine's and indolamines at 264 nm and the native fluorescence emission collected and recorded using a UV-enhanced CCD array (Spec-10; 2KBUV/LN; Princeton Instruments; Trenton, NJ, USA). The CE separations were performed by hydrodynamic injection of 10 nL of sample and using a BGE of 25 mM citric acid (pH 2.5, applied voltage +30 kV) or 50 mM borate (pH 9.5, applied voltage +21 kV). Analytes were identified based on comparison of both the migration time and fluorescence spectrum to that of standard mixtures of analytes.

CE-ESI-MS analysis was performed using a Bruker Microtof or a Maxis (Bruker Daltonics; Billerica, MA, USA) mass spectrometer for detection. All separations were performed using 1% formic acid in water as the electrolyte and applied voltage of +30 kV. The sheath liquid was 0.1% formic acid in 50/50 methanol/water. Samples were hydrodynamically injected for a total volume of ~ 6 nL. Mass spectra were collected and recorded at a rate of 2 Hz with calibration was performed using sodium formate clusters. Analytes were identified based on comparison of both the CE migration time and mass match to that of standard mixtures of analytes.

SM17. 2 **Chemicals, Reagents and Animals**

All chemicals for buffers were purchased from Sigma-Aldrich, and standard amino acids were purchased from Fluka. Ultrapure Milli-Q water (Milli-Q filtration system, Millipore, Bedford, MA) was used for all buffers, standard solutions, and sample preparations.

Ctenophores *Pleurobrachia and Beroe* were collected at the University of Washington Friday Harbor Laboratories (FHL), San Juan Island, WA. *Mnemiopsis leidy* were collected at Marine Biological Laboratory (Wood Hole. MA) and at the Whitney Laboratory (University of Florida, St. Augustine area, FL) during spring-summer of 2008-2012. Specific body parts of ctenophores were dissected under a stereomicroscope using microscissor and tweezers. The samples were immediately stored in a 0.5mL Eppendorff tube containing Milli-Q water at -80 ^{o}C until use.

SM 17.3 **Amino Acids Microanalysis using CE with LIF**

We used the same laser-induced fluorescence set up and CE assays as described elsewhere[65]. Briefly, CE coupled with the ZETALIF detector (Picometrics, France) was used for the assay of amino acids and selected transmitters (see **Supplementary Tables 23S-25S**). In this work a helium−cadmium laser (325nm) from Melles Griot, Inc. (Omnichrome® Series56, CARLSBAD, CA) was used as the excitation source. Before the photomultiplier tube (PMT), the fluorescence was both spatially filtered using a machined 3-mm pinhole and wavelength filtered. All instrumentation, counting, and high-voltage CE power supply were controlled using a DAx 7.3 software.

All solutions were prepared with ultrapure Milli-Q water (Milli-Q filtration system, Millipore, Bedford, MA) to minimize the presence of impurities. 30 mM borate buffer (pH 9.5) was used for sample preparation. All solutions were filtered using 0.2-μm filters to remove particulates. Buffers were degassed by ultrasonication for 10 min to minimize the chance of bubble formation. In order to detect non-fluorescent analytes, they were derivatized with OPA; specifically, 75mM OPA/β-mercaptoethanol (ME) stock solution was prepared by dissolving 10mg of OPA in 100μL of methanol and mixing with 1mL of 30mM borate and 10μL of β-ME. 10mM stock solution of amino acids and neurotransmitters were prepared by dissolving each compound in 30mM borate. OPA and β-ME were stored in a refrigerator, and fresh solutions were made weekly.

All experiments were conducted using a 75cm length of 50μm I.D. × 360μm O.D. fused silica capillary (Polymicro Technologies, AZ). 30mM borate and sodium dodecyl sulfate (SDS)

electrolyte (adjusted to pH 9.5 with NaOH) was used for a separation buffer of amino acids analysis. For separation steps, the capillary inner-wall was successively washed with 1M NaOH, Milli Q water, and the separation buffer by applying pressure (1900mbar) to the inlet vial. Then a sample was loaded using electrokinetic injection (8kV for 12s) following a derivatization with *o*-phthaldialdehyde (OPA). The separation was performed under a stable 20kV voltage at 20˚C.

SM17. 4 **Data Analysis**

Once an electropherogram was acquired, peaks were confirmed by both electrophoretic mobility of each analyte and spiking corresponding standards into the sample. Five-point calibration curves (peak area vs. concentration) of analytes were constructed for quantification using standard solutions, and the dilution factor is required to be considered for an original concentration. The 3σ method is used to determine limit of detection that is calculated from standard deviations of blank (n=5) and the calibration slope m of low concentration standards. The reproducibility and accuracy of the method were evaluated by calculating relative standard deviation (RSD) and error values. In order to obtain the peak area, a baseline was constructed and subtracted using a derlim algorithm of DAx software version 7.3 (Van Mierlo Software Consultancy, the Netherlands). A statistical data analysis was performed by Sigma Plot software (SPSS, Inc., Richmond, CA).

Supplementary Tables

Supplementary Table 1S

Summary of genomic DNA libraries constructed from *Pleurobrachia bachei*

454			Illumina		
Library Type	**# Animals**	# of Plates All *Titanium*	**Library Type**	**# Animals**	# Sequencing lanes
Fragmented (Shotgun)	1 Animal (2009)	3 Full plates	Paired end	1 Animal (2010)	3 lanes GA, 2 lanes HiSeq 2000 4 flowcell MiSeq
3Kb Long-Tag Paired End	10 animals (2009)	1 Full plate	Mate pair 3-4Kb insert	10 animals (2010)	1 lane GA
5Kb Long-Tag Paired End	20 animals (2009)	1 Full plate	Mate pair 4-5 Kb insert	Pool 2 Animals (2010)	1 lane GA 2 lanes HiSeq 2000
			Mate pair 4-6 Kb insert	Pool 2 Animals (2010)	1 lane HiSeq 2000
			Mate pair 6-9 Kb insert	Pool 2 Animals (2010)	1 lane HiSeq 2000

Supplementary Table 2S

Summary of *Pleurobrachia bachei* genome sequencing

Library type/ # of Animals	**454**			**Illumina: MiSeq/HiSeq2000**		
	Reads	**Bases**	**SRA: Project/ Sample**	**Reads**	**Bases**	**SRA: Project/ Sample**
Shotgun 1 animal	3,286,344	1,177,702,794	SRP001155			
Paired end 1 animal for Illumina HiSeq / pool for 454	2,145,046	746,499,725	SRP001155	401,143,926	38,358,073,560	SRP001155
Paired end 1 animal for MiSeq (250x2)				138,900,000	32,700,000,000	
Mate pair Pool of Animals HiSeq				612,457,003	59,033,324,028	SRP001155
TOTAL	**5,431,390**	**1,924,202,519**		**1,152,500,929**	**130,091,397,588**	

TOTAL Genomic Sequencing: 132,015,600,107 bp or ~ **132 Gb** or 733-825x physical coverage of the *Pleurobrachia* genome (the size of the *P. bachei* genome is estimated as ~160-180 Mb).

See details in the **Supplementary Methods SM1-2.**

Supplementary Table 3S

Summary *Pleurobrachia bachei* transcriptome sequencing (All data are located in SRP000992, NCBI)

Tissue	454		Illumina		Ion torrent	
	Reads	**Mbp**	**Reads**	**Mbp**	**Reads**	**Mbp**
Adult Tissues from						
P. bachei						
Aboral Organ	177,462	18.7				
Tentacles	468,850	120.1				
Combs	371,399	88.9				
Whole animal	17,162	1.8	236,203,645	23,240.6		
Embryos, mixed stages	241,517	48.2	228,616,963	22,259.0		
Small RNA sequencing			21,549,716	775.7		
Mouth					2,811,271	335.62
Stomach					2,280,134	314.26
Developmental Stages						
P. bachei						
1-cell					128,124	11.39
2-cells					113,375	11.7
4-cell					270,046	25.27
8-cell					1,601,384	183.64
16-cell					211,751	17.59
32-cell					197,499	15.95
64-cell					1,592,020	146.68
Gastrulation					276,141	43.76
1-Day					304,684	25.4
3-Day					2,443,666	233.25
Total	1,276,390	277.8	486,370,324	46,275.5	12,230,095	1,364.51
TOTAL	**499,699,347 Reads**			**47,917 Mbp or ~47.9 Gbp**		
Other Ctenophore Species						
Beroe abyssicola			45,444,644	4,544,464,400		
Bolinopsis infundibulum			43,377,170	4,337,717,000		
Coeloplana astericola			41,685,220	4,168,522,000		
Dyrodora glandiformis			41,269,166	4,126,916,600		
Euplokamis dunlapae			68,302,698	6,830,269,800		
Mnemiopsis leidyi			49,782,900	4,978,290,000		
Pleurobrachia pileus			50,626,422	5,062,642,200		
Mertensiidae sp			47,454,246	4,745,424,600		
Vallicula multiformis			49,090,372	4,909,037,200		
TOTAL			**437,032,838**	**43,703 Mbp**		

The *Pleurobrachia bachei* RNA-seq assemblies and Comparative Ctenophora transcriptome assemblies as well as their annotations are accessible at: http://moroz.hpc.ufl.edu/

See details in the **Supplementary Methods SM1-2.**

Supplementary Table 4S

Pleurobrachia bachei transcriptome assembly: Summary

Assembly	Input Size (sequences)	# Assembly Bases	Number Contigs	Number Singletons	Average Length	Max Length	*Number of unique transcripts annotated
Tentacles[1]	468,850	5,975,854	2,790	29,274	307bp	2256bp	8,120
Combs[1]	371,399	30,267,075	40,382	82,989	245bp	2442bp	21,041
Aboral Organ[1]	177,462	4,208,916	1,637	23,798	335bp	1476bp	4,331
Combined Adult[1]	1,017,711	45,487,515	48,671	130,410	367bp	2502bp	128,695
Adult[2]	236,203,645	54,263,467	202,636	1	268bp	6,888bp	79,750
Embryos Mix[2]	228,616,963	17,829,586	69,101	1	257bp	4,335bp	-
1-cell embryos[3]	128,124	8,043,989	1,106	91,662	86bp	403bp	3,642
2-cell embryos[3]	113,375	3,159,508	462	46,641	67bp	391bp	4,848
4-cell embryos[3]	270,046	9,251,419	8,773	108,936	78bp	634bp	15,702
8-cell embryos[3]	1,601,384	54,591,558	35,757	559,242	91bp	751bp	11,944
16-cell embryos[3]	211,751	7,789,479	5,385	116,352	63bp	638bp	11,699
32-cell embryos[3]	197,499	7,287,894	5,126	113,443	61bp	577bp	11,385
64-cell embryos[3]	1,592,020	22,863,442	22,105	355,256	60bp	713bp	35,841
Gastrulation[4]	276,141	13,209,999	10,426	822	136bp	850bp	4,617
1 day embryos[4]	304,684	5,445,541	10,305	457	137bp	1,316bp	30,458
3 day embryos[4]	5,544,489	23,793,838	166,163	12,506	133bp	980bp	14,888

[1] Two-stage Newbler/Paracell Transcriptome Assembler pipeline
[2] Velvet
[3] CLCBio *de novo* assembly tool
[4] MIRA
Sequencing Technologies are as following: blue 454, green Illumina, black Ion Torrent
* Number of transcripts with at least one alignment to the UniProtKB/Swiss-Prot protein database.

Support of Gene Models in Transcriptomes

Non-redundant gene models were aligned to our assembled 454 and Illumina transcriptomes with tblastn. 18,994 (85%) gene models were supported by the 454 transcriptome assembly and 15,954 (71%) were supported by the Illumina transcriptome assemblies. In this instance, support of a gene model is considered to be at least one alignment of an assembled transcript to the model with greater than 90% identity over the length of the alignment. Taking into account the union of the two alignment results, a total of 19,996 (89%) of the predicted gene models obtained some level of support. Due to our extensive *Pleurobrachia* transcriptome sequencing (~1,790x coverage), we removed any predicted gene model lacking such support. A further 463 partial gene models were removed, by merging them with another gene model using transcript sequence as a scaffold. This brought the final count of *Pleurobrachia* gene models to **19,523**.

See details in the **Supplementary Methods SM2-4.**

Supplementary Table 5S

***Pleurobrachia bachei* Draft Genome Assembly v1.1**

Estimated genome size	~160 Mb
Number of Bases within Scaffolds (>2Kbp)	145,572,285 bp
Number of Scaffolds(>2Kbp)	12,102
Average Scaffold Length (>2Kbp)	12,028 bp
Max Scaffold Length(>2Kbp)	320,313 bp
Scaffold N50 (>2Kbp)	1,407
Scaffold N50 Length (>2Kbp)	23,607 bp
Scaffold N90 (>2Kbp)	7,325
Scaffold N90 Length(>2Kbp)	4,701 bp
Protein coding Gene Models	**19,523**
GC Content in Genome	35.6%
Average Exon Length	211 bp
Average Number of Exons per Gene	5
Percent of Genome in Exons	17.5%
Average gene size	3,750 bp
GC Content in Exons	45.7%
Percent of Genome in Coding Sequence (exon/intron)	17.5% / 52.4%
Median Intron size	186 bp
Average Intron size	508 bp
Percentage of Genome in Introns	34.9%
Pleurobrachia-Specific Genes* (NR/SP)	10, 897/11,957 (56%/61.2%)
Gene Homologs to Metazoa* (NR/SP)	8,626/7,566 (44%/38.8%)

*based on BLAST of the gene models to NR and SwissProt, with an evalue cutoff of 1e-04 (NR: 8,626 shared, 10,897unique; SwissProt: 7,566 shared, 11,957 unique)

The *Pleurobrachia* genome assembly is available at: http://moroz.hpc.ufl.edu/

Summary: The average *Pleurobrachia* gene contains 5 exons, 4 introns, and is 3,750 bp long, having an average exon length of 211bp.

Supplementary Table 6S

Dinucleotide content in the *Pleurobrachia bachei* genome

Dinucleotide	Count	Frequency*
AA	11,414,571	0.073112
TT	11,381,757	0.072902
AT	10,394,429	0.066578
TG	9,206,734	0.058970
CG*	**3,909,499**	**0.025041**
GC*	**4,795,633**	**0.030717**
AG	9,086,093	0.058198
GA	9,194,097	0.058890
CA	9,206,079	0.058966
TC	9,207,216	0.058974
TA	9,295,437	0.059539
CT	9,102,348	0.058302
AC	8,215,256	0.052620
GT	8,212,475	0.052602
CC	7,104,208	0.045503
GG	7,102,207	0.045491
NC	4052	2.6E-05
CN	4032	2.58E-05
AN	4557	2.92E-05
GN	4181	2.68E-05
NG	4103	2.63E-05
NT	4530	2.9E-05
TN	4457	2.85E-05
NA	4542	2.91E-05

*relatively lower occurrence of CpGs dinucleotides in genomes suggest the presence of DNA methylation[66] (see details in the **Supplementary Data SD3**).

Supplementary Table 7S

Homology of *Pleurobrachia bachei* gene models (19,523) to public databases

	NCBI-NR Database				SwissProt Database			
E-value Cutoff	Shared Genes with Eukaryotic Organisms		Ctenophore Specific Genes		Shared Genes with Eukaryotic Organisms		Ctenophore Specific Genes	
1e-10	6,876	35.2 %	12,648	64.8 %	5,869	30.1 %	13,655	69.9%
1e-04	8,626	44.2 %	10,897	55.8 %	7,566	38.8 %	11,958	61.2%
1	8,693	44.5 %	10,831	55.5 %	7,645	39.2 %	11,879	60.8%

The gene models used were predicted out of the genome assembly that had all prokaryotic contamination removed (**Supplementary Methods SM2.3**). BLASTp was conducted under default conditions and a word size of 3 against the public protein databases of NCBI-NR and UniProtKB/SwissProt.

Supplementary Table 8S

Comparison of *Pleurobrachia bachei* draft genome with other assembled animal genomes

	Pleurobrachia bachei* NGS	*Amphimedon queenslandica*[11] - **S**	*Acropora digitifera*[67]** NGS + **S**	*Hydra magnipapillata*[12] **S**	*Nematostella vectensis*[68] **S**	*Trichoplax adhaerens*[69] **S**	*Ascaris suum*[70] - *	*Crassostrea gigas*[71]** NGS + **S**	*Strongylocentrotus purpuratus*[72] **S**
Estimated genome size	160 Mbp	190 Mbp	420 Mbp	1.3 Gbp	450 Mbp	98 Mbp	309Mbp	545-637Mbp	814 Mbp
Assembly Summary									
#Scaffolds (>2Kb)	12,102	13,522	4,765	37,311	10,804	1,415	NA	5,079	NA
# Bases within Scaffolds	145 Mbp	167 Mbp	419 Mbp	1,256 Mbp	297 Mbp	98M bp	272 Mbp	558 Mbp	1.019 Mbp
Scaffold N50 (Scaffolds)	1,407	310	642	3,526	181	6	1,618	401	18,000
Scaffold N50 Length (Scaffolds)	23,607 bp	120,000 bp	191,500 bp	84,.400 bp	420,000bp	5.79 Mbp	408 kbp contigs	401,319 bp	65,601 bp
Scaffold N90 (Scaffolds)	7,325	3,492	2,241	38,825	NA	NA	748 contigs	1,669	NA
Scaffold N90 Length (Scaffolds)	4,701 bp	5,000 bp	50,000 bp	1,924 bp	NA	NA	80,017 bp contigs	67,587	NA
GC Content in Genome	35.6 %	NA	39.0 %	29.0 %	NA	NA	37.9 %	NA	36.9%
Transposable elements	8.5 %	NA	25.4 %	NA	26.2 %	NA	NA	36.1%	NA
% Polymorphism	1.3(6.5)***%	2.3 %	NA	1 %	0.8%	1 %	NA	1.3% WT	NA
Annotation Summary									
Gene Models	19,523	>30,000	23,668	31,452	27,273	11,520	18,542	28,027	23,300
Transcriptome support of Gene Models	96 %	83 %	78 %	NA	NA	NA		96%	NA
Average gene size	3,750 bp	NA	7,600 bp	NA	4,500 bp	1375bp	6,536 bp	NA	7,700 bp
Average Exon Length	211 bp	NA	230 bp	NA	208 bp	163 bp	153 bp	NA	NA
Average Intron size	508 bp	80 bp	NA	NA	NA	285 bp	1,081 bp	NA	NA
Average Number of Exons per Gene	5	5	7	NA	5.3	8.4	6	NA	NA

NA - Not available
* Sequenced with only next-generation sequencing technology
** Sequenced with a combination of next-gen and Sanger
S - sequenced with Sanger

*** genome average vs max observed in a selected gene set (in parentences)

Supplementary Table 9S

Content of Transposable Elements (TEs) in *Pleurobrachia bachei* and ***Nematostella vectensis* genomes**

Classes of TEs	***P. bachei* %**	***N. vectensis* %**
Total DNA transposons	**5.1**	**18.5**
"cut and paste":		
Mariner (Pogo group)	0.13	2.3
hAT	0.16	2.1
Kolobok	0.09	1.6
PiggyBac	0.002	1.0
Harbinger	0.07	1.0
Academ	0.01	0.01
Transib/Chapaev	0.05	0.01
Sola	0.01	0.02
MuDR	0.03	0.3
En/Spm	0.01	0.05
Merlin	0.01	0.01
ISL2EU	0.001	0.002
P	-	0.5
Crypton	-	0.001
Unclassified non-autonomous	3.78	5.2
"self-synthesizing" *Polintons*	0.34	3.0
"rolling circle" *Helitrons*	0.36	1.4
Total retrotransposons	**2.6**	**4.9**
LTR retrotransposons:	**0.3**	**2.0**
Gypsy	0.28	1.5
Copia	0.04	0.05
BEL	-	0.2
Unclassified	-	0.2
DIRS	-	**0.4**
Non-LTR retrotransposons:	**2.3**	**2.5**
CR1 (CR1, L2, Daphne,Rex1)	0.73	1.0
RTE (RTE, RTEX)	0.19	0.4
Cre	0.32	0.01
Hero	0.04	-
Loa, Ingi	0.39	-
R2 (R2, R4)	0.11	0.01
L1 (L1, Tx1)	-	0.1
SINEs (SINE2)	0.07	0.3
Unclassified	0.23	-
Penelope	0.19	0.7
Unclassified TEs	**0.9**	**2.8**
Total TEs	**8.5**	**26.2**

See details in **Supplementary Data SD2**

Supplementary Table 10S

Comparative Ctenophora Transcriptomes: Assembly summary

Assembly	Input Size (sequences)	# Assembly Bases	Number Contigs	Avg. Length	Max Length	*Number Annotated
Beroe abyssicola	45,444,644	39,028,800	39,349	991bp	12,764bp	16,805
Bolinopsis infundibulum	49,090,372	119,979,450	159,357	752bp	11,233bp	43,700
Coeloplana astericola	53,379,682	73,388,845	144,134	509bp	9,270bp	29,508
Dyrodora glandiformis	41,685,220	37,664,309	47,526	792bp	8,124bp	10,946
Euplokamis dunlapae	41,269,166	94,177,040	128,177	734bp	10,772bp	34,600
Mnemiopsis leidyi	68,302,698	117,237,257	164,117	714bp	8,934bp	70,150
Pleurobrachia sp.	43,377,170	76,480,245	114,898	665bp	8,978bp	99,854
Mertensiidae sp.	47,454,246	83,648,543	111,307	751bp	10,674bp	31,609
Vallicula multiformis	49,924,286	58,757,388	147,015	399bp	9,173bp	7,321

All sequencing performed with IlluminaHiSeq
Assemblies performed with the Trinity *de novo* transcriptome assembler.
* Number of transcripts with at least one alignment to the UniProtKB/Swiss-Prot protein database.

Supplementary Table 11S

Comparative Transcriptomics in Ctenophora: Number of Recognized Homologs* of *Pleurobrachia* Gene Models in Transcriptome Assemblies from Different Ctenophore Species

RNA-seq Assemblies from different Ctenophore Species	**19,523 Filtered *Pleurobrachia* Gene Models****	
	8,626 annotated genes	**10,897 unannotated genes**
Pleurobrachia pileus	8,669	8,138
Beroe abyssicola	7,936	2,573
Bolinopsis infundibulum	8,046	3,091
Coeloplana astericola	7,774	4,064
Dyrodora glandiformis	7,471	2,231
Euplokamis dunlapae	7,803	2,080
***Mertensiidae* sp.**	7,761	2,841
Vallicula multiformis	8,134	3,011
Mnemiopsis leidyi (RNA-seq)	8,345	3,362
Mnemiopsis leidyi (gene models)	8,414	3,840
	12,254	
Aligned Homologs	**8,634 (99.7%)**	**8,413 (78%)**
TOTAL Homologs	**17,147 (78%)**	

*****Across all of the comparisons, a total of 19,523 filtered *Pleurobrachia* gene models were aligned to at least of one other ctenophore species. In summary, 17,147 of all *Pleurobrachia* genes, or ~78%, were aligned to another ctenophore. Interestingly, of 8,626 annotated filtered gene models, 8,634 of them (or ~99%) had an alignment to one other ctenophore; in contrast, only 8,413 of the 10,897 unannotated gene models (~78%) were aligned to at least one other ctenophore species. All alignments were done with blastp for protein-protein alignments and tblastn for protein-rna alignments with an e-value cutoff of 1e-04. Finally, ~ 63% of *Pleurobrachia* gene models (or 12,254) were aligned to *Mnemiopsis* gene models (16,548 including 10,183 annotated).

** of 32,207 unfiltered gene models (17,717 annotated and 14,490 unannotated), 25,747, including 8,683 unannotated, were aligned to other ctenophores.

Supplementary Table 12aS

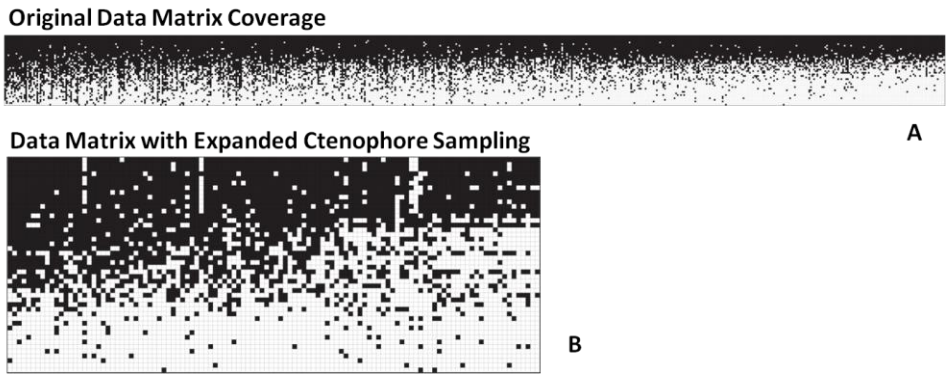
Initial Sampling of Species for Phylogenomic Analysis of Animal Relationships

Phylum	O.T.U.	Species	Type	Source	Sampled Genes	Leaf Stability
Fungi	*Spizellomyces*	*Spizellomyces punctatus*	Sanger	dbEST	95	0.9608
	Piromyces	*Piromyces* sp.	Sanger	dbEST	179	0.9608
	Glomus	*Glomus intraradices*	Sanger	dbEST	275	0.9605
	Blastocladiella	*Blastocladiella emersonii*	Sanger	dbEST	248	0.9605
	Saccharomyces	*Saccharomyces cerevisiae*	Genome	InParanoid	567	0.9611
Ichthyosporea	*Amoebidium*	*Amoebidium parasiticum*	Sanger	dbEST	36	0.9597
	Capsaspora	*Capsaspora owczarzaki*	Sanger	dbEST	94	0.9574
	Sphaeroforma	*Sphaeroforma arctica*	Sanger	dbEST	116	0.9597
Choanoflagellata	*Monosiga brevicollis*	*Monosiga brevicollis*	Genome	JGI	515	0.9620
	Monosiga ovata	*Monosiga ovata*	Sanger	dbEST	315	0.9620
Placozoa	*Trichoplax*	*Trichoplax adhaerans*	Genome	JGI	554	0.8485
Cnidaria	*Acropora*	*Acropora millepora*	Sanger	dbEST	304	0.9560
		Acropora palmata	Sanger	dbEST		
	Aiptasia	*Aiptasia pallida*	Sanger	dbEST	91	0.9560
		Aiptasia pulchella	Sanger	dbEST		
	Anemonia	*Anemonia viridis*	Sanger	dbEST	183	0.9560
	Clytia	*Clytia hemisphaerica*	Sanger	dbEST	333	0.9549
	Hydra	*Hydra vulgaris*	Sanger	dbEST	481	0.9549
	Hydractinia	*Hydractinia echinata*	Sanger	dbEST	63	0.9554
	Nematostella	*Nematostella vectensis*	Genome	JGI	562	0.9560
	Podocoryne	*Podocoryne carnea*	Sanger	dbEST	43	0.9554
	Porites	*Porites astreoides*	Sanger	dbEST	56	0.9560
Porifera	*Amphimedon*	*Amphimedon queenslandica*	Sanger	dbEST	218	0.8989
	Aphrocallistes	*Aphrocallistes vastus*	454	This study	124	0.7888
	Carteriospongia	*Carteriospongia foliascens*	Sanger	dbEST	29	0.8987
	Ephydatia	*Ephydatia muelleri*	Sanger	dbEST	56	0.8989
	Heterochone	*Heterochone calyx*	Sanger	dbEST	102	0.7888
	Oopsacas	*Oopsacas minuta*	Sanger	dbEST	22	0.7888
	Oscarella	*Oscarella carmela*	Sanger	dbEST	172	0.8904
		Oscarella lobularis	Sanger	dbEST		
	Leucetta	*Leucetta chagosensis*	Sanger	dbEST	52	0.8367
	Sycon	*Sycon coacturm*	454	This study	47	0.8367
		Sycon raphanus	Sanger	dbEST		
Ctenophora	*Pleurobrachia*	*Pleurobrachia bachei*	454	This study	586	0.9546
		Pleurobrachia bachei	Genome	This study		
	Pleurobrachia	*Pleurobrachia pileus*	454	dbEST	409	0.9546
	Mnemiopsis	*Mnemiopsis leidyi* (dbEST)	Sanger	dbEST	135	0.9092
		Mnemiopsis leidyi (Dunn et al. 2008)	Sanger	Trace		

						0.9553
Deuterostomia	*Branchiostoma*	*Branchiostoma floridae*	Genome	JGI	564	
	Ciona	*Ciona intestinalis*	Genome	InParanoid	580	0.9553
	Homo	*Homo sapiens*	Genome	InParanoid	586	0.9553
	Saccoglossus	*Saccoglossus kowalevskii*	Sanger	dbEST	405	0.9553
	Strongylocentrotus	*Strongylocentrotus purpuratus*	Genome	InParanoid	530	0.9553
Lophotrochozoa	*Aplysia*	*Aplysia californica*	UniGene	UniGene	171	0.9548
	Lottia	*Lottia gigantea*	Genome	JGI	565	0.9548
	Capitella	*Capitella teleta*	Genome	JGI	570	0.9548
Ecdysozoa	*Caenorhabditis*	*Caenorhabditis elegans*	Genome	InParanoid	584	0.7998
	Daphnia	*Daphnia pulex*	Genome	JGI	572	0.9548
	Drosophila	*Drosophila melanogaster*	Genome	InParanoid	586	0.9548
	Ixodes	*Ixodes scapularis*	Sanger	dbEST	455	0.9548

See details in the **Supplementary Methods SM7** and **Supplementary Data SD4**

Supplementary Table 12bS. Phylogenomic data matrix coverage.



Original data matrix of 586 genes (**A, Supplementary Table 12aS**), and for data matrix with expanded ctenophore sampling of 114 genes (**B**, **Supplementary Table 13S**, see below). Genes are ordered along the X-axis from left to right from best sampled to worst sampled. Taxa are ordered along the Y-axis from top to bottom from most genes sampled to fewest genes sampled. Black squares represent a sampled gene fragment and white squares represent a missing gene fragment.

Supplementary Table 13S

Expanded Sampling of Species for Phylogenomic Analysis of Animal Relationships

Species	Type	Reads	Source	Sampled Genes
Monosiga brevicollis	Genome	-	JGI	94
Monosiga ovata	Sanger	29,495	dbEST	38
Salpingoeca rosetta	Genome	-	Broad	97
Trichoplax adhaerans	Genome	-	JGI	101
Acropora digitifera	Genome	-	marinegenomics.oist.jp	66
Aiptasia pallida	Sanger	10,295	dbEST	14
Anemonia viridis	Sanger	39,939	dbEST	23
Clytia hemisphaerica	Sanger	85,991	dbEST	37
Gorgonia ventalina	Illumina	36,941,899	SRA	105
Hydractinia echinata	Sanger	9,460	dbEST	7
Hydra vulgaris	Sanger	167,982	dbEST	57
Nematostella vectensis	Genome	-	JGI	100
Podocoryne carnea	Sanger	3,466	dbEST	4
Porites astreoides	Sanger	11,516	dbEST	5
Amphimedon queenslandica	Sanger	63,542	dbEST	30
Aphrocallistes vastus	454	224,948	This study	9
Carteriospongia foliascens	Sanger	3,556	dbEST	5
Ephydatia muelleri	Sanger	1,696	dbEST	7
Heterochone calyx	Sanger	4,790	dbEST	6
Leucetta chagosensis	Sanger	5,703	dbEST	4
Oopsacas minuta	Sanger	1,443	dbEST	6
Oscarella carmela	Sanger	11,176	dbEST	8
Sycon coactum	454	204,257	This study	8
Beroe abyssicola	Illumina	45,444,644	This study	72
Bolinopsis infundibulum	Illumina	43,377,170	This study	52
Coeloplana astericola	Illumina	41,685,220	This study	48
Dryodora glandiformis	Illumina	41,269,166	This study	69
Euplokamis dunlapae	Illumina	68,302,698	This study	53
Mertensiidae sp.	Illumina	47,454,246	This study	56
Mnemiopsis leidyi	Illumina	49,782,900	This study	71
Pleurobrachia pileus	Illumina	50,626,422	This study	86
Pleurobrachia bachei	Genome	80,151	This study	60
Pleurobrachia sp.	Sanger	35,290	dbEST	37
Vallicula sp.	Illumina	49,090,372	This study	60
Branchiostoma floridae	Genome	-	JGI	108
Ciona intestinalis	Genome	-	InParanoid	103
Homo sapiens	Genome	-	InParanoid	108
Saccoglossus kowalevskii	Sanger	202,190	dbEST	47
Strongylocentrotus purpuratus	Genome	-	InParanoid	104
Aplysia californica	Sanger	-	UniGene	15
Capitella teleta	Genome	-	JGI	103
Crassostrea gigas	Genome	-	http://gigadb.org/pacific_oyster/	108
Lottia gigantea	Genome	-	JGI	106
Daphnia pulex	Genome	-	JGI	105
Drosophila melanogaster	Genome	-	InParanoid	108
Ixodes scapularis	Sanger	193,773	dbEST	64

Supplementary Table 14aS

Data sources for the analysis of gene gain and gene loss across Metazoa

Species	Source URL	**Input data size (MB)**
Arabidopsis thaliana	ftp://ftp.arabidopsis.org/home/tair/home/tair/ webcite	14.72
Capsaspora owczarzaki	http:/ / www.broadinstitute.org/ annotation/ genome/ multicellularity_project/ GenomeDescriptions.html webcite	6.2
Monosiga brevicollis	http://genome.jgi-psf.org/Monbr1/Monbr1.home.html webcite	5.77
Amphimedon queenslandica	http://getentry.ddbj.nig.ac.jp/top-e.html?ACUQ00000000 webcite	11.87
Pleurobrachia bachei	The present study	
Trichoplax adhaerens	http://genome.jgi-psf.org/Triad1/Triad1.home.html webcite	14.72
Nematostella vectensis	http://genome.jgi-psf.org/Nemve1/Nemve1.home.html webcite	9.76
Daphnia pulex	http://wfleabase.org/genome/Daphnia_pulex/ webcite	10.94
Lottia gigantea	http://genome.jgi-psf.org/Lotgi1/Lotgi1.home.html webcite	9.72
Strongylocentrotus purpuratus	http://www.spbase.org/SpBase/download/ webcite	22.07
Saccoglossus kowalevskii	ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Skowalevskii/fasta webcite	8,14
Homo sapiens	http://www.gencodegenes.org/ webcite	42.55
Branchiostoma floridae	http://genome.jgi-psf.org/Brafl1/Brafl1.home.html webcite	24.34

Supplementary Table 14bS. Uniquely shared and lineage-specific genes among basal metazoans

Values under species names indicate the number of genes (*****) without any recognized homologs (*e* value is 10^{-4}) vs the total number of predicted gene models (**bold**) in are relevant species.

	Pleurobrachia	***Amphimedon***	***Trichoplax***	***Nematostella***
	10,899*	12,050*	18,702*	15,257*
Pleurobrachia	**19,523**			
Amphimedon	6,083	**30,060**		
Trichoplax	2,894	2,297	**27,416**	
Nematostella	7,208	10,537	5,803	**27,273**
Pleurobrachia/Amphimedon/ Nematostella	5,777			
Pleurobrachia/Trichoplax/ Nematostella	2,591			
Pleurobrachia/Amphimedon/ Trichoplax/Nematostella	2,142			

Supplementary Table 15S

Reduced Metazoan Gene Complements in Ctenophores

	Choanozoa *Monosiga*	**Ctenophora *Pleurobrachia***	Porifera *Amphimedon*	Placozoa *Trichoplax*	Cnidaria *Nematostella*	Bilateria *Homo*
Hox	-	-	-	1	9	40
Hedgehog	-	-	-	-	2	3
Cys-loop R	-	-	-	-	21	60
Churchill	-	-	1	1	1	1
TRBP2	-	-	1	-	1	1
Drosha+Pasha	-	-	2	1	2	2
#of miRNAs	-	-	8	-	49	1,424
Immune Genes						
RDH	-	-	1	-	2	5
MyD88	-	-	1	-	1	1
Toll-like R	-	-	2	-	5	11
Nod-like R	-	-	61	12	72	20
Transmitters						
Glutamate	-	+	-	-	+	+
GABA	-	-	-	-	+	+
Dopamine	-	-	-	-	+	+
Octopamine	-	-	-	-	-	+
Serotonin	-	-	-	-	+	+
Acetylcholine	-	-	-	-	+	+
Nitric Oxide	-	-	+	+	+	+

Illustrative examples of genes primarily absent in Ctenophores (*Pleurobrachia*) as compared to other sequenced genomes. Nitric oxide – NOS gene is absent. Note the lack of majority of neurotransmitter systems in *Pleurobrachia* as well as immune, miRNA and HOX complement.

Supplementary Table 16S

Examples of Gene Gain and Ctenophore Novelties

	Choanozoa *Monosiga*	**Ctenophora *Pleurobrachia***	Porifera *Amphimedon*	Placozoa *Trichoplax*	Cnidaria *Nematostella*	Bilateria *Homo*
RNA editing	-	**9**	2	1	1	3
Nova	-	**6**	-	1	1	2
Collagen IV	-	**7**	-	-	2	6
Gap Junctions	-	**12**	-	-	-	27
Glu-signaling						
iGluR		**14**	-	11	5	18
Glu Transporters	2 (Sialins)	**8**	2	1	3	1
Glutaminases	5	**8**	1	1	2	2
ASIC	-	**29**	18	2	25	9
Voltage-gated K channels	-	**27**	1	11	50	40
Neurexin	-	**1**	-	-	-	3
Divergent SIX Homeodomain	-	**11**	-	-	-	-
Novel Secretory Peptides	-	**72**	-	-	-	-
Ctenophore Specific Genes	-	**10,831**	-	-	-	-

Illustrative examples of gene gain and the ctenophore-specific expansion of selected gene families. See the main text for details.

Supplementary Table 17S

Metazoan Transcription Factors

Transcription Factor	Filasterea	Choanoflagellata	Ctenophora	Porifera	Placozoa	Cnidaria	Bilateria
	Capsaspora	***Monosiga***	***Pleurobrachia***	***Amphimedon***	***Trichoplax***	***Nematostella***	***Homo***
bHLH	31	11	26	18	26	72	69
bZIP	25	15	15	20	16	35	38
HMG box	9	12	19	19	19	36	45
MADS-box	2	3	2	2	3	3	4
Homeodomain	9	2	92	32	36	153	162
HOX	-	-	-	-	**1**	**9**	**40**
Forkhead	4	11	22	20	18	33	31
CP2	2	1	2	3	1	2	4
Churchill	1	-	-	1	1	1	1
T-Box	4	-	5	10	5	13	16
RUNT	2	-	2	1	1	1	2
p53 DNA binding	**1**	**2**	**1**	**1**	**1**	3	3
STAT DNA binding	1	1	1	2	6	1	3
RDH	1	-	-	1	-	2	5
MH1+/ MH2 (Smad)	-	-	6	9	4	5	8
Ets	-	-	11	8	7	16	18
Nuclear Hormone Receptor	-	-	2	2	4	17	70

Notes:
bHLH – basic-helix-loop-helix;

HOX genes are absent in Ctenophora, Porifera, Choanoflagellata and Filasterea

Churchill or Rel/NF-kappaB (RDH) genes are absent in *Pleurobrachia*

There are also a notable expansion of 92 homeodoman, basic Helix-Loop-Helix (bHLH) and Forkhead TFs relative to sponges and placozoans

See details in **Supplementary Data SD5.2** . All *Pleurobrachia* sequences used in the analysis are listed in **Supplementary Excel Table 31S**.

Supplementary Table 18S

The Homeodomain gene complement in the *Pleurobrachia bachei* genome

Class	Subclass	Number of HB genes	HB gene families in *Pleurobrachia* genome
ANTP	HOXL	0	Absent
	NKL	48	Msx, Hmx, NK6, NK7.1, BarX, NK2.2/2.8, Hex
PRD	PAX	1[a]	PAX2/5/8
	PRDL	13	Mix, Otp, Ceh-10, Unc-4, Otx
LIM		5	Lim3, Lmx, Lim1, Islet, Apterous
POU		3	Pou1, Pou4, Pou6
SIX		14	Six1/2, Six3/6
TALE		4	Pbc, Meis, Prep
HNF		1	
CUT		1[b]	One Cut
ZF		1	Zeb
CERS		1	
PROS		0	
		Total=92	

Supplementary Table 19S

miRNA pathways in Metazoa

		Dicer	Ago	TRBP2	Piwi	Hen1	**Drosha**	**Pasha**	# miRNAs
Bilateria	*Homo*	1	4	1	4	1	1	1	1424
	Drosophila	2	2	2	3	1	1	1	238
	Caenorhabditis	1	5	1	3	1	1	1	207
	Capitella	1	1	1	2	1	1	1	125
	Lottia	1	1	1	2	1	1	1	60
Cnidaria	*Nematostella*	2	3	1	3	1	1	1	89
Placozoa	*Trichoplax*	5	1	-	-	-	1	-	-
Porifera	*Amphimedon*	4	2	1	3	2	1	1	8
Ctenophora	*Pleurobrachia*	1	3	-	2	1	-	-	-
Choanoflagellata	*Monosiga*	-	-	-	-	-	-	-	-

Several classes of small RNAs have been discovered including small interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs) and microRNAs (miRNAs); each of them has distinctive molecular targets and biogenic pathways with key enzymes presented in this table. In contrast to other representative metazoan genomes, *Pleurobrachia* has the most reduced complement of small RNA processing enzymes (lacking *Drosha*, *Pasha*, and TRBP2) suggesting the absence of the canonical the miRNA pathways. These predictions are consistent with (I) our phylogenetic analyses of *Dicer* and *Aro* genes and (II) absence of recognized miRNAs (see details in the Supplementary Data **SD5.4**). All *Pleurobrachia bachei* sequences used in this table can be found in **Supplementary Excel Table 31S** and the cloned genes were submitted to Gene Bank (*Dicer* JN202325, *Ago*2 JN798505, *Ago*3 JN798506, *Ago*4 JN798507, *Piwi*1 JN798508, *Piwi*2 JN798509) these were also used for *in situ* hybridization tests.

Supplementary Table 20aS

Pattern recognition receptors in metazoans

		SRCR[1]	**NLR**	**TLR**	**MyD88**	PGRP
Deuterostomia	*Homo sapiens*	81(16)	20	11	1	4
	Strongylocentrotus purpuratus	1,095(218)	203	253	4	5
Ecdysozoa	*Drosophila melanogaster*	14(7)	0	9	1	13
	Caenorhabditis elegans	3(1)	0	1	0	0
Lophotrochozoa	*Capitella capitata*[2]	110(67)	0	105	1	7
	Helobdella robusta[2]	35(20)	0	16	1	0
Cnidaria	*Nematostella vectensis*	134(67)	72	5	1	0
	Hydra magnipapillata	23(6)	290	1[3]	1	0
Placozoa	*Trichoplax adherens*	143(46)	12	0	0	1
Porifera	*Amphimedon queenslandica*	1,422(300)	61	2[3]	1	0
Ctenophora	***Pleurobrachia bachei***	**137(56)**	**0[4]**	**0**	**0**	**2**
Choanoflagellates	*Monosiga brevicollis*	3(1)	0[5]	0	0	0

[1] Numbers of SRCR domains in the genome are shown. Numbers in parentheses indicate the total number of gene models.
[2] See Davidson et al.[73]
[3] These species possess divergent, TLR-related sequences (see[74,75]).
[4] We identified five NACHT domains in *Pleurobrachia bachei*, but there was no evidence for other NLR domains nearby. These may represent divergent NLR sequences.
[5] Although NLRs have not been described in *Monosiga brevicollis*, we identified NACHT domains in this genome.

Supplementary Table 20b: Genes related to immunity in *P. bachei*

Gene	Gene model
PGRP	scaffold330.1|size62058 #FGENESH: 6
	scaffold1526.1|size21519 #FGENESH: 1
Mif	scaffold607.1|size44186 #FGENESH: 8
	scaffold867.1|size34511 #FGENESH: 9
	scaffold21056.1|size761 #FGENESH: 1
MACPF	scaffold5481.1|size5700 #FGENESH: 1
	scaffold1926.1|size17301 #FGENESH: 2
	scaffold7237.1|size4303 #FGENESH: 1
	scaffold2637.1|size12745 #FGENESH: 2
	scaffold2455.1|size13751 #FGENESH: 4
	scaffold2482.1|size13585 #FGENESH: 3
	scaffold1094.1|size28902 #FGENESH: 6
	scaffold1134.1|size28133 #FGENESH: 5
	scaffold13259.1|size1421 #FGENESH: 1
	scaffold145.1|size92176 #FGENESH: 20
	scaffold1331.1|size24211 #FGENESH: 4
IRF	scaffold3370.1|size9205 #FGENESH: 1
	scaffold382.1|size57346 #FGENESH: 3
	scaffold6998.1|size4591 #FGENESH: 1
	scaffold7836.1|size3571 #FGENESH: 1
ETS	scaffold1204.1|size26785 #FGENESH: 6
	scaffold1212.1|size26574 #FGENESH: 4
	scaffold12642.1|size1526 #FGENESH: 1
	scaffold1766.1|size18797 #FGENESH: 4
	scaffold2400.1|size14013 #FGENESH: 2
	scaffold285.1|size66915 #FGENESH: 8
	scaffold329.1|size62070 #FGENESH: 7
	scaffold329.1|size62070 #FGENESH: 7
	scaffold4060.1|size8358 #FGENESH: 1
	scaffold4423.1|size7341 #FGENESH: 1
	scaffold536.1|size47761 #FGENESH: 1
	scaffold614.1|size44031 #FGENESH: 3
	scaffold7467.1|size3989 #FGENESH: 1
Gata	scaffold557.1|size46459 #FGENESH: 2

Supplementary Table 21S

The distribution of RNA binding proteins across phyla

RNA-binding PFAM Domains	*Monosiga brevicollis*	***Pleurobrachia bachei***	*Amphimedon queenslandica*	*Trichoplax adhaerens*	*Nematostella vectensis*	*Drosophila melanogaster*	*Mus musculus*
RRM-1	75	**80**	121	85	125	117	227
KH	9	**31**	18	14	20	27	42
Dead Box	38	**41**	72	60	54	63	>100
DsRM	3	**11**	10	10	13	8	21
Other	27	**96**	45	45	45	41	>100
Total	**152**	**260**	**266**	**203**	**257**	**256**	**>490**

Adapted from Kerner et al., 2011 [76]

Supplementary Table 22S

Microchemical Determination of Canonical Low Molecular Weight Neurotransmitters in *Pleurobrachia* using Capillary Electrophoresis (CE) coupled with Laser-Induced Native Fluorescence (LINF) and Electrospray Mass Spectrometry Detection Schemes

Analyte	CE-LINF		CE-ESI-MS		Notes
	Detection Limit (nM)	**Result in** *Pleurobrachia*	**Detection Limit (nM)**	**Result in** *Pleurobrachia*	
Serotonin	1	ND	NA	--	All tissues
Tryptophan	5	D	<125 nM	D	All tissues
Tyrosine	88	D	<125 nM	D	All tissues
Dopamine	36	ND	<125 nM	ND	All tissues
Norepinephrine	91	ND	<125 nM	ND	All tissues
Epinephrine	94	ND	<125 nM	ND	All tissues
Octopamine	10	ND	<125 nM	ND	All tissues
L-Dopa	NA	--	<125 nM	ND	All tissues
5-Hydroxy Tryptophan	NA	--	<125 nM	ND	All tissues
Tryptamine	2	ND	<125 nM	ND	All tissues
Indole 3 Acetic acid		ND	NA	--	All tissues
Indole butyric acid		ND	NA	--	All tissues
Histamine	NA	--	<125 nM	ND	All tissues
Acetocholine	NA	--	<125 nM	ND	All tissues
Tyramine	25	ND	<125 nM	ND	All tissues
Serine	NA	--	<125 nM	D	In Combs
GABA	NA	--	<125 nM	D	In Combs
Aspartate	NA	--	<125 nM	D	In Combs
Glutamate	NA	--	<125 nM	D	In Combs

D – detected, ND - non-detected, NA – detection limit is <100 nM, n=4

Supplementary Table 23S

Correlation Coefficients (r) **and Limits of Detection** (**LOD**) **for selected Amino Acids and Neurotransmitters:** Capillary Electrophoresis (CE) coupled with Laser-Induced Fluorescence the ZETALIF detector (Picometrics, France), n=4

Analyte	**r**	**LOD** (nM)
Citrulline	0.9985	149
Serine	0.9963	62
5-hydroxy tryptophan	0.9961	73
Tyramine	0.9992	10.7
GABA	0.9991	76.1
Glycine	0.9923	128
Taurine	0.9953	95.8
Phenylalanine	0.9982	5.57
Dopamine	0.9853	69.2
Tryptophan	0.9959	7.14
Glutamate	0.9962	22.3
Aspartate	0.9972	16.5
Arginine	0.9674	113
Octopamine	0.9562	48
Serotonin (5-HT)	0.9213	85

Supplementary Table 24S

Metabolite Concentrations in Ctenophores (Capillary electrophoresis assays)

Species	Region	Cit	Ser	GABA	Gly	Tau	Phe	Trp	Glu	Asp	Arg
Beroe abyssicola	AO[a]	25.0±1.6	2.6±0.3	0.22±0.01	2.25±0.18	0.61±0.07	1.97±0.06	2.21±0.13	9.94±0.83	3.78±0.2	1.88±0.15
	Mouth	38.7±2.1	4.46±0.45	0.43±0.04	7.47±0.90	2.49±0.21	4.88±0.38	5.29±0.37	9.45±0.49	6.74±0.58	9.93±1.37
	Comb	24.8±1.06	2.91±0.17	0.91±0.08	6.53±0.38	0.89±0.08	3.51±0.3	4.76±0.22	11.2±0.42	4.40±0.28	1.50±0.07
	Pigment Spots	22.1±0.73	2.03±0.05	2.71±0.17	7.85±0.54	0.58±0.01	11.2±0.3	9.14±0.18	4.69±0.17	3.08±0.01	10.8±0.25
Pleurobrachia bachei	Comb	5.99±0.1	1.26±0.1	0.34±0.02	0.91±0.07	23.7±0.5	33.7±1	1.63±0.07	5.49±0.2	1.58±0.08	27.4±1.3
	Body wall	2.82±0.1	0.75±0.08	0.23±0.02	0.28±0.03	30.5±2.6	2.86±	1.50	4.17±0.2	2.30±0.07	1.58±0.1
	AO[a]	8.81±0.3	1.23±0.1	1.48±0.1	6.84±0.7	22.8±0.5	10.9±0.04	1.73±0.04	10.4±0.7	3.49±0.01	31.4±4.2
	Mouth	15.7±1.5	1.57±0.1	0.42±0.04	4.24±0.3	20.6±1.7	4.07±0.3	3.75±0.3	6.39±0.6	2.08±0.2	3.32±0.1
	Stomach	44.0±3.2	4.95±0.3	2.81±0.2	15.6±1.1	64.3±4.2	12.1±0.8	17.7±0.9	25.4±1.6	7.40±0.6	13.4±1.0
	Tentacle	24.5±1.9	3.16±0.2	0.77±0.08	1.38±0.07	20.4±1.6	1.40±0.1	1.94±0.2	8.20±0.4	6.66±0.6	12.0±1.4

a. WB (whole body), AO (aboral organ)
Results are in μM±SD, n=5

Abbreviations: Cit – citrulline, Ser – serine, Gly – glycine, Tau – taurine, Phe - phenylalanine; Trp – tryptophan, Glu - glutamate, Asp – aspartate, Agr – arginine.

Supplementary Table 25S

L- and D - Glutamate and Aspartate Concentrations in Ctenophores
(Capillary electrophoresis assays)

Species	Region	L-Glu	D-Glu	D-Asp	L-Asp
Beroe abyssicola	AO[a]	25.5±1.9	2.48±0.5		9.74±0.8
	Mouth	9.54±1.9	0.67±0.1		6.41±1.6
	Combs	14.7±3.1	1.59±0.09		7.53±1.3
Pleurobrachia bachei	Combs	27.1±2.03			9.66±0.71
	Body wall	10.1±3.39			10.2±5.39
	AO[a]	6.54±0.61	0.41		7.45±0.32
	Mouth	12.3±1.36	1.39		11.2±0.85
	Stomach	36.8±3.27	3.14±0.37		16.9±1.04
	Tentacle	11.7±0.65	2.72±0.32		9.82±0.57

a. WB (whole body), AO (aboral organ)
Results are in μM±SD, n=5

Abbreviations: L-Glu - L-glutamate; D-Glu - D-glutamate; L-Asp – L-aspartate; D-Asp – D-aspartate.

Supplementary Table 26aS

26aS Neurotransmitter Signaling in Metazoan

	Ctenophora	**Porifera**	**Placozoa**	**Cnidaria**	Chordata	
	Pleurobrachia	*Amphimedon*	*Trichoplax*	*Nematostella*	*Ciona*	*Homo*
Glutamatergic Signaling						
AMPA Receptor	-	-	-	-	1	4
Kainate Receptor	-	-	-	-	1	5
Delta Receptor	-	-	-	-	2	2
NMDA Receptor	-	-	-	3	3	7
Other Ionotropic GluR	**14**	-	12	>9	4	-
Metabotropic GluR	3	>20	9	>8	8	8
Cholinergic Signaling						
Nicotinic(α,β,δ,ε,γ) Receptor	-	-	-	>21	14	17
Muscarinic Receptor	-	-	-	-	5	5
Choline acetyltransferase	-	-	-	1?	1	1
GABAergic Signaling						
$GABA_A$(α,β,δ,ε,π,θ)Receptor	-		-	>11	5	19
$GABA_B$ Receptor	-	>20	16	4	1	2
Glutamate decarboxylase	1	1	1	1	1	2
Glycine Signaling						
GlycineA Receptor	-	-	-	1	4	4
GlycineB Receptor	-	-	-	-	-	1
Histamine Signaling						
Histamine Receptor 1-4	2	-	-	?	1	4
Histamine decarboxylase	-	-	-	-	1	1
Serotonergic Signaling						
Serotonin 3A-E	-	-	-	-	-	5
Serotonin 1,2,4-7	NA	NA	NA	NA	>5	13
Tryptophan hydroxylase	-	-	-	-	1	2
Catecholamine Signaling						
Dopamine (1,2) Receptor	NA	NA	NA	NA	3	5
Adrenergic(α,β) Receptor	NA	NA	NA	NA	6	9
TH/AADC/DH	-/-/-	-/-/1	-/-/1	-/1/1	1/1/1	2/1/1
ATP gated channels						
P2X	1	2	2	2	0	7
Amiloride sensitive channels						
ASIC-like	29	18	3	25	27	4
ENaC						5

NA: Not applicable to distinguish between receptor subtypes.

26bS Neurotransmitter Signaling in Metazoa: 7 Transmembrane receptor family

7 Transmembrane Receptor Family	**Ctenophora** *Pleurobrachia*	**Porifera** *Amphimedon*	**Placozoa** *Trichoplax*	**Cnidaria** *Nematostella*	**Chordata** *Ciona*	*Homo*
7TMR1 (rhodopsin-like)	223	145	343	826	126	683
7TMR2 (Secretin/adhesion)	16	70	26	32	17	48
7TMR3 (sweet-taste-like)	11	44			8	26
7TMR (glutamate-like)			49	34	6	22
7TMR (frizzled-like)	2		2	6	5	11
7TMR (cAMP-type)			1		1	
Other		3	3	3		8
Total	**252**	**259**	**424**	**901**	**163**	**798**

Supplementary Table 27S

Number of Ionotropic Glutamate Receptors in Metazoans

Species	**Total**	AMPA	Kainate	NMDA			Delta	Other	Plant
				type1	type2	type3			
Homo sapiens	**18**	4	5	1	4	2	2		
Ciona intestinalis	**11**	1	1	1	1		2	5	
Branchiostoma floridae	**10**	3	3					4	
Strongylocentrotus purpuratus	**10**	2	2					6	
Saccoglossus kowalevskii	**11**	1	1	1	1	2	1	4	
Drosophila melanogaster	**14**	2	10	1	1				
Caenorhabditis elegans	**15**	3	4	1	1		4	2	
Aplysia californica	**13**			1	1			11	
Lottia gigantea	**12**			1	1			10	
Nematostella vectensis	**10**			1				9	
Trichoplax adhaerens	**12**	0	0	0	0	0		12	0
Amphimedon queenslandica	**0**	0	0	0	0	0		0	0
*Pleurobrachia bachei**	**14**	0	0	0	0	0		14	
Arabidopsis thaliana	**20**								20

*Ionotropic glutamate receptors can not be classified in terms of AMPA/Kainate/NMDA subtypes (see details in **Supplementary Data SD5.9**)

Supplementary Table 28S: Parameters for Prediction of Secretory Peptides*

Prediction Stringency	SignalP NN	SignalP HMM	TargetP	Expected of cleavage sites	Phobius TM	Phobius SP
Least	90%>x≥50%	90%>x≥50%	x=1	x>0	No	Yes
Medium	90%>x≥50%	90%>x≥50%	x=1	x>1	No	Yes
Most	x≥90%	x≥90%	x=1	x>1	No	Yes

SignalP NN and HMM are the probabilities produced by the SignalP Neural Network and Hidden Markov model Respectively. TargetP is the certainty score (1-5) produced by TargetP. Phobius TM is Phobius' prediction of a transmembrane domain. Phobius SP is Phobius' prediction of a signal peptide. Finally, the number of proteins matching these prediction criteria.

* In order to achieve an open but tractable initial prediction set in the absence of prior studies, we used different cutoff criteria and developed three non-overlapping sets of secretory peptide models. Here, we combined multiple distinct heuristic tools in our secretory peptide prediction pipeline, to filter and rank potential secretory products from both genome and transcriptome data. Because almost all pre-secreted peptides are localized to endoplasmic reticulum via presence of an N-terminal signal peptide consisting of 5-20 hydrophobic residues, this heuristic forms the strongest marker of secreted products. For signal peptide prediction, we employed the SignalP[27] and TargetP[28] tools. SignalP utilizes separate Hidden Markov Model (HMM) and Neural Network (NN) components, trained on a variety of organisms, for the sole purpose of secretory signal peptide prediction. TargetP utilizes the Neural Network component of SignalP and separate Neural Networks to discriminate between 'secreted', 'mitochondrial', and 'other' localization signals. To eliminate 'contamination' with putative transmembrane proteins, TMHMM[29] is employed to predict transmembrane topology. Transmembrane domains, like signal peptides, are characterized by a region of hydrophobic residues of similar length, often causing false-positive signal peptide or transmembrane results from trained prediction methods such as TMHMM. Phobius[30] uses complementary HMM, trained separately for transmembrane domains and signal peptides, to call a consensus.

The difference in final predictions between simply removing TargetP criteria, and setting the certainty cutoff to 2 was minimal, due to overlap with SignalP. Additionally, initial tests revealed that changing the target certainly cutoff from 1 to 2 nearly doubled the positive prediction count uniformly across other variables. Thus, it was kept at 1 for all selection sets. Phobis compares the probability scores of transmembrane predictions and signal peptide predictions directly, and thus does not produce its own probability score; hence, adherence to 'signal peptide with no transmembrane domain' is strictly enforced for all predictions. We observed signal cutoff to have the greatest effect on prediction numbers. The default signal and target certainly cutoff of 50% for both NN and HMM proved too lenient for our purposes.

Supplementary Excel Tables (included as separate items)

Description of **Supplementary Excel Tables 29-33**

Excel Table 29S *Pleurobrachia bachei* Unfiltered Gene Models

Obtaining *P. bachei* unfiltered gene models was described in **Supplementary Methods,** section **S3.1.** There were 32,207 predicted gene models before removing redundancy. All transcripts have an annotation note. Included is the database ID number, scaffold location of the prediction (filtered gene models only), and the predicted amino acid sequence. The best annotation is the lowest e-value obtained either from National Center of Biotechnology Information's (NCBI) non-redundant protein database (NR) or the UniProtKB/*Swiss-Prot* (SP) protein database. Also included, if applicable, KEGG ID with associated KEGG pathway, PFAM ID with associated description and Gene Ontology (GO) ID with associated description S**upplementary Methods,** section **S4.1.** Finally, a relative abundance (quantification) of all the *P. bachei* RNA-seq experiments is incorporated for all major embryonic and developmental stages (1-cell, 2-cells, 4-cells, 8-cells, 16-cells, 32-cells, 64-cells, early and later gastrula, 1-day, 3-days larvae) as well all major adult tissues and organs (combs, mouth, tentacles, stomach, the aboral organ). Relative level of expression (quantification) is reported as a normalized frequency where the number of reads that hit to a specific transcript is divided by the total number of reads in a particular sequencing project. Actual numbers of all reads for each RNA-seq project are reported in Table S3.

Excel Table 30S *Pleurobrachia bachei* Filtered Gene Models

Obtaining *P. bachei* unfiltered gene models was described in **Supplementary Methods**, section **S3.1.** After removing redundancy, there were 19,523 predicted gene models. All transcripts have annotation notes as described above.

Excel Table 31S Analyzed Transcripts

Any *Pleurobrachia bachei* sequence described or used for any analysis including phylogenetic analysis or expression profiling is located here. Include is the same as above except for no expression data is presented on the subset of transcripts. There are 10 sheets in this excel file and they include the following list of Excel sheets in the Table 29S:

1. Analyzed transcripts
   DNA methylation transcripts
   siRNA core transcripts
   RNA editing transcripts
   Collagen transcripts
   Developmental transcripts
2. All RNA binding proteins transcripts
3. All Transcription factor transcripts
4. Homeobox domain transcripts
5. Ion channel transcripts
6. Ionotropic glutamate receptor transcripts
7. ASIC channel transcripts
8. Neuronal transcripts
9. Innexin transcripts
10. Processing enzymes
11. G-protein coupled receptors
12. Pleurobrachia specific transcripts
13. Immunity transcripts

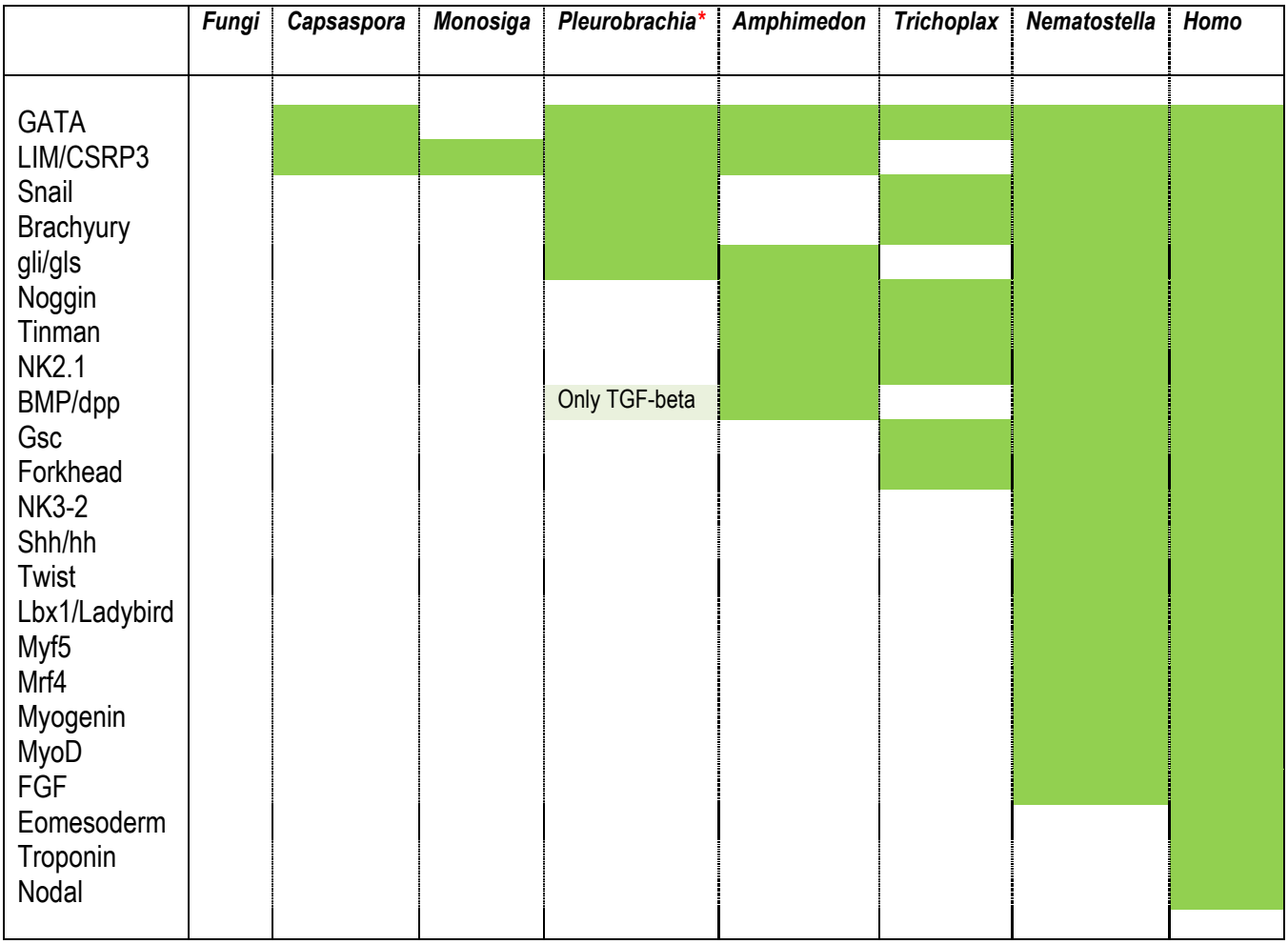
Excel Table 32S Predicted Secretory Peptides

There were 72 prohormone/precursors for predicted secretory peptides. All are presented with their associated NCBI accession number, name, DNA and protein sequence.

Excel Tables 33. Divergence of Ctenophore Proteins

Supplementary Table 34S: Presence/Absence of orthologs of genes encoding pre- and post-synaptic proteins in the *Pleurobrachia* genome

Supplementary Table 35S: Presence/Absence of muscle/mesoderm specification genes in the *Pleurobrachia* genome



Supplementary Data

The Genome of the North Pacific Ctenophore, *Pleurobrachia bachei*

Supplementary Data: Table of Contents

SD1. Biology of *Pleurobrachia bachei* (A. Agassiz, 1860)**: Background information about the organism**

Comb jellies[77] are beautiful holoplanktonic predators (**Supplementary Videos**) with highly sophisticated ciliated locomotion mediated by specialized comb plates (**Fig. 1d**, main text), giving them the name *Cteno-phora*, meaning comb-bearing. Ctenophores have unique food capture mechanisms mediated by glue-secreting cells – colloblasts (**Fig. 1e**). Complex behaviours are controlled by atypical neurons forming two distinct nets with unconventional synapses and a well-differentiated muscular apparatus including muscle-derived conductive and integrative systems[78,79]. Highly deterministic development[77,78] includes ctenophore-specific gastrulation, mesoderm specification and biradial symmetry with tentacle and mouth axes. Little is known about the molecular mechanisms of development[80] and nervous organization in Ctenophora[78,79,81].

Pleurobrachia bachei is a pelagic animal (**Fig. 1a**, main text, and **Extended Data Fig. 1a-e, Supplementary Videos**) that lives, often in swarms, near the surface of the coastal northeast Pacific Ocean. It represents a classical ctenophore body plan with preservation of all canonical Ctenophora[78] features such as cydippid larva, well-developed ciliary locomotion and tentacles – some or all of these features are absent in other ctenophore lineages,see below. As such, the cydippid *Pleurobrachia* is considered a very important model for comparative analysis both within the ctenophore clade and to characterize evolutionary trends and relationships at the larger scale among animal lineages.

The body of *P. bachei* is nearly spherical, to about 2 cm diameter, and is soft and jellylike, of glassy transparency. The outside of this spherical body possesses eight evenly-spaced longitudinal ctene rows, each about 1 mm wide and tapering at each end; the ctene rows run nearly from pole to pole. Each of the ctene rows is composed of about 50 small, overlapping, colorless combs with beautiful iridescence patterns (**Extended Data Fig. 1c**), each comb is composed of hundreds of long, compound cilia[82] (**Fig. 1e)**, which beat together, thus propelling the animal. Normally, the effective stroke of a comb plate is directed towards the aboral end, so that the ctenophore typically swims mouth first[83]. The combs beat sequentially, passing metachronal waves along ctene rows in the opposite direction to the effective stroke, and the comb rows are separately controlled so that the animal can swim forward, or backward, and even spin[84].

At one end (the "**Oral pole**"), the eight comb rows converge near a cruciform mouth. At the other end (the "**Aboral pole**") is a complex ciliary structure that functions as an 'elementary brain', coordinating and organizing the beating of the comb rows as well as feeding and escape responses, and sensing orientation of the entire animal using a calcareous statolith supported by balancer cilia[79].

The nervous system otherwise is composed of two diffuse nerve nets (**Fig. 5a**, main text): One is just below the body surface, the sub-epithelial net (**Extended Data Fig. 1h**), and the second nerve net is located in the mesogleal tissues and probably has a mesodermal origin. Apparently, the second nervous system has no obvious homologs or analogs in any other groups of basal metazoans or bilaterians studied so far[85].

Internally, there is a broad, but flat, pharynx-gut extending below the mouth to about the center of the animal (**Extended Data Fig. 1b)**, in which food is digested, and a system of narrow, transparent canals which run under each of the eight comb rows and connect to the pharynx-gut near the equator of the animal. The aboral canal extends from the base of the pharynx to the aboral pole, where it terminates in two small excretory (aboral) pores. Nutrients are distributed throughout these canals; waste passes out the excretory pores or may also be ejected through the mouth.

P. bachei is a carnivorous predator, capturing small zooplankton prey on two light pink, contractile, filamentous tentacles, each of which has numerous fine side branches called tentilla (**Extended Data Fig. 1a** and **4a**). Tentacles exit through two small holes on opposite sides of the body closer to the aboral pole and can extend out many cm beyond the animal, or can be contracted entirely into two sheaths inside the body that angle up towards the gut[86].

Typical feeding behavior for *P. bachei* is to swim upward (mouth facing up) in a broad arc or spiral, deploying its tentacles outward in a large spiral web, using a combination of muscle relaxation and friction drag in the water. Once tentacles are fully set, with side branches also extended, the ctenophore stops swimming and waits quietly as an "ambush predator"[87] until prey encounter the large filamentous web or until the ctenophore changes location. Each tentacle is covered with hundreds of thousands of microscopic sticky cells called colloblasts (**Fig. 1e** and **Extended Data Fig. 4a**), which will adhere to and eventually subdue prey that impacts the web[88]. When contacted by prey, the tentacle contracts and the body spins to bring the food-laden tentacle in proximity of the mouth. This is accomplished by reversing the beat of the comb rows on the side of the capture tentacle. The tentacle and its prey will eventually pass by the mouth and the prey is released as it is ingested in the pharynx[84]. Each *P. bachei* sets its tentacular net through several liters of sea-water, so in water where populations of this ctenophore are numerous, the small zooplankton may be almost completely strained from the water. Most prey eaten by *P. bachei* are less than a few mm in size, consisting mostly of small crustaceans, primarily copepods and cladocera and crustacean eggs and larval stages[89].

Unlike most ctenophores, all species of *Pleurobrachia* lack the ability to produce any kind of bioluminescence[90].

SD1.2 **Reproduction**

Pleurobrachia bachei, like most ctenophores, is a simultaneous hermaphrodite, meaning that a single individual can produce both eggs and sperm at the same time. Gametes are free-spawned into the surrounding ocean through a series of pairs of sperm ducts and oviducts that exit through pores in the epidermis above each of the eight meridional canals, between each pair of comb plates[91]. In damaged specimens including some animals subjected to careful handling in the laboratory, some of the gametes may enter the meridional canals and exit through either the mouth or anal pores, but this is not normal[91,92]. Sperm are released a few minutes prior to egg release and both Dunlap[91] and Hirota[93] report that *P. bachei* is self-fertile; the eggs have a thin and tough, transparent vitelline membrane that is further protected by a jelly coat, received in the

oviduct. Fertilization and subsequent development occur freely in the sea without any parental protection. All major embryonic and developmental stages are summarized in the **Fig 4a)**.

Pleurobrachia bachei, is a direct developer like all ctenophores, the embryo develops directly into a tiny larva (cydippid larva) with the same body plan of the adult. Animals spawn in the early morning, and it appears that they respond to both a circadian and lunar cycle (B.J. Swalla personal observations). Early cleavages are fast, with cleavage occurring about every 30 minutes after the one cell stage, even at 10°C. Cydippid larvae develop within about 10 hours, and the ctene rows appear and begin beating, even though the larvae are not yet hatched. Stereotypical cleavages result in a group of large vegetal M macromeres with the small E micromeres on the animal half. Gastrulation begins at about 4 hours after fertilization with epiboly of the micromeres over macromeres and invagination of the stomodeum.

In well-fed *P. bachei*, spawning of up to about 1000 eggs, with a daily average closer to 350 eggs, and uncounted numbers of sperm will occur many days in succession[93]. It is estimated that the natural *P. bachei* population off La Jolla, California, has a doubling time of about 35 days[89] and that early production of small numbers of gametes by 1-2 mm diameter "post-larvae" contributes an important fraction of the total reproduction of the population[89,93].

SD1.3 **Range and Habitat**

Pleurobrachia bachei is known from the Pacific coast of North America from at least Southeast Alaska (Sitka) to Acapulco, Mexico[94]. Off Southern California, where a comprehensive population study has been carried out, *P. bachei* is generally confined to surface waters (the upper 50 m), with most individuals in the upper 15 m during the day and at about 30 m at night, with the densest populations within 5 km of shore[89]. In Saanich Inlet, British Columbia, *P. bachei* resides from the surface to 110 m over a 200 m bottom, with higher numbers collected from the upper 25 m; it can be found throughout the 210 m deep water column in the nearby and more thoroughly mixed waters of Stuart Channel, B.C.[95,96]. The population of *P. bachei* in Friday Harbor, Washington, where *P. bachei* was collected for this genome study, seems to be renewed probably several times annually by large numbers of animals drawn from the Pacific Ocean through the Strait of Juan de Fuca and into the Salish Sea (a system of inland marine deep channels and basins that includes Friday Harbor) (Mills, unpublished data). *P. bachei* ctenophores are sometimes found washed up on NE Pacific beaches, where they look like soft, transparent, marbles above the surf line.

SD1.4 **Etymology**

The genus name comes from the Greek *pleura*, side and *brachion*, arm. The species is named for Professor Alexander Dallas Bache, superintendent of the Coast Survey and friend of Louis Agassiz. Bache hired Louis Agassiz' son, Alexander Agassiz, as an engineer on a ship surveying the United States/Canada boundary in the Strait of Georgia between Washington and British Columbia, where this species was first collected in the summer of 1859; it was also found at the entrance to Admiralty Inlet that summer[97-99].

Pleurobrachia is a genus of cydippid ctenophore, which is globular, or nearly spherical in shape, bearing a pair of branched tentacles that emerge from opposite sides, and is one of the most common coastal genera of ctenophores in the world. The common name sometimes used for this kind of ctenophore is "**sea gooseberry**."

SD1.5 **How many species of *Pleurobrachia?***

More than ten species of *Pleurobrachia* have been described, but the most commonly recognized species in zoological literature are *Pleurobrachia pileus*[85], *Pleurobrachia rhodopis*[100] and *Pleurobrachia bachei*[98]. Taxonomy of all of species of *Pleurobrachia* is in need of careful review and revision, probably best done using molecular/genetic techniques, but for the moment, we will briefly summarize the differentiation of these three most-studied species. Discussion of the current status of cydippid taxonomy can be found elsewhere[89,90,92-95,101-117] and the history can be found in Bigelow (1912) who made the last, most thorough comparison of these three species, having studied preserved material and illustrations of living material from three oceanic locations[102] .

Briefly, the ctenophore *Pleurobrachia pileus* is recognized as the name of choice for animals on both sides of the North Atlantic. *Pleurobrachia rhodopis* is described as a species endemic to the Mediterranean, and *Pleurobrachia bachei* occurs along the west coast of North America (see **Range and Habitat**, above).

Elsewhere in the Pacific, ctenophores studied live without preservation artifacts have been previously identified as *P. pileus* in Hawaii[89], *P. pileus*, in New Zealand[104,118], as *Pleurobrachia* sp. in Australia by Gershwin *et al.*, 2010, as *P. pileus* in nearshore Chile[106] and South Atlantic[112,119] as well and as *P. bachei* over Chilean seamounts[117].

Scientists studying the ctenophore fauna off the southern French Mediterranean coast, including Hernandez-Nicaise[120], Franc[121], Laval[122] , Buecher and Gasser[111], Molinero[116], and others have consistently used the name *Pleurobrachia rhodopis*. Recent molecular studies using ctenophores collected from the same location near Villefranche-sur-Mer now identify these animals as *P. pileus*[85,123-126]. Scientists working in the Black Sea typically assign the name *P. rhodopis*[127,128] although some have acknowledged that *P. rhodopis* is a synonym of *P. pileus*[109,129].

One to seven additional described species of *Pleurobrachia* appear to be valid[102,108]. Other globose, transparent cydippid ctenophores are assigned to the genera *Hormiphora*, which is characterized partially by having a pair of relatively long tentacle bases that lie parallel to, and close to the walls of the pharynx.

Our own sampling and initial phylogenomic analysis performed on two distinct populations of *Pleurobrachia* collected from Central and Northern Atlantic respectively (**Fig. 2**) suggests that they belong to the same species *P. pileus*; and this species morphologically and molecularly is distinct from the second Pacific *Pleurobrachia* species which we identified as *P. bachei*. Thus, the species ***Pleurobrachia bachei*** collected in the north eastern Pacific at the Friday Harbor (WA) has been sequenced in the current study.

SD2 Features of Transposable Elements (TE) in the *Pleurobrachi bachei* Genome

The anthozoan cnidarian *Nematostella vectensis* is the only species among basal metazoans where TEs have been systematically identified and studied[68]. Thus, the *Pleurobrachia* genome provides important insights for these elusive and dynamic genome elements. In the *Pleurobrachia* genome, most families of autonomous TEs (**Supplementary Table_9S**) coding for transposases and DNA polymerases (DNA transposons) and reverse transcriptase (non-LTR and LTR retrotransposons) are composed of a few, often incomplete, copies. Therefore, identification of complete sequences of these TEs, including non-coding termini, was challenging. In such cases, these TEs have been described by their coding regions. Using WU-BLAST/CENSOR, we identified all genomic fragments of the genome similar to the consensus sequences that were considered as copies of TEs. Second, given these consensus sequences, we found automatically all insertions longer than 30-bp present in the identified copies of the protein-coding TEs. Insertions have been treated as putative novel TEs. They have been clustered based on their pairwise DNA identities and replaced by consensus sequences built for each cluster. After manual refinements of the consensus sequences, identified families of TEs were classified based on their structural hallmarks, including target site duplications, terminal repeats, encoded proteins and similarities to TEs classified previously[130,131]. The same approach, combining BLAST-based searches and *de-novo* detections was applied recently in several genome projects[131]. Identified TEs have been named according to the standard nomenclature and are deposited in Repbase. As a result, we identified more than 300 families of TEs, **Supplementary Table 9S** which occupy at least 8.5% of the *P. bachei* genome with numerous examples of diversification of some ancient TE classes (e.g. coding for transposases, reverse transcriptases and transposon-specific proteins) suggesting that the lineage leading to *Pleurobrachia* experienced extensive evolutionary changes.

In general, the *Pleurobrachia* genome is characterized by similarly high diversity of TEs (**Supplementary Table 9S**) but a reduced representation of some TE families. Contrary to *Nematostella*, the *Pleurobrachia* genome does not contain *P* and *Crypton* DNA transposons, *BEL* LTR retrotransposons, *L1* non-LTR retrotransposons and *DIRS* retrotransposons. However it contains *Loa*, *Ingi* and *Hero* non-LTR retrotransposons, which are lacking in the anemone, *Nematostella*. Most families of ctenophore autonomous TEs that code for transposases and DNA polymerases (DNA transposons) and reverse transcriptase (non-LTR and LTR retrotransposons) are composed of only one or a few copies. On the other hand, some families of non-autonomous DNA transposons were quite successful in terms of proliferation. For instance, the *DNA-8-2*_PBa family, the most numerous family among all identified TEs, is represented by more than 7000 copies. All diverse families of TEs found in the *Pleurobrachia* are "young": elements from the oldest families are less than 10% divergent from their consensus sequences and their ORFs coding for transposases, reverse transcriptases, and other transposon-specific proteins are only slightly damaged by mutations. On the contrary, the *Nematostella* genome contains much "older" TEs: oldest elements >25% divergent from their consensus sequences[68]. Such difference implies that frequent and relatively long deletions were behind quick elimination of TEs in the ctenophore genome. By this the ctenophore genome is similar to the compact *Arabidopsis thaliana* and *Drosophila melanogaster* genomes[132,133].

In terms of their bulk contribution to the genome size, ctenophore DNA transposons are twice more abundant than retrotransposons (**Supplementary Table 9S**. Remarkably, some non-autonomous ctenophore DNA transposons that form several families and subfamilies are the shortest DNA transposons identified so far in eukaryotes. These are *DNA-8-3F*_PBa (40-bp long), *DNA-8-3E*_PBa (41-bp), *DNA-8-6A*_PBa (49-bp), *DNA-9-5C*_PBa (53-bp), *DNA-8-7*_PBa (55-bp). All these transposons have short imperfect terminal inverted repeats (5'-TTG and CAA-3' termini), with 8 or 9 bp target site duplications. Although its autonomous counterparts remain to be found, it is most likely that they form a novel superfamily of DNA transposons. In summary, analysis of TEs suggests that the lineage leading to *Pleurobrachia* was a subject of extensive and, possible recent evolutionary changes.

SD3 Analysis of DNA methylation in *Pleurobrachia*

DNA methylation is found across many domains of life including bacteria, plants, and animals. It is essential for normal development and maintenance of cell phenotypes. As such DNA methylation is the quintessential epigenetic mark and, probably, the oldest of epigenetic mechanisms. In prokaryotes, DNA methylation occurs on both cytosine and adenine bases and is part of the host restriction modification system[134]. In plants, the methylation of cytosines occurs throughout all sequence contexts and is involved in gene silencing as well as - unique transposon silencing[135]. In animals, DNA methylation is predominantly confined to cytosines in the dinucleotide pair (CpG), thus creating a mechanism for cellular memory. The idea that the pattern of unmethylated and methylated CpG was copied when cells divide was first proposed in 1975[136,137]. The consequence of this mechanism is that methylation patterns are inherited following certain types of cell divisions, and DNA methylation patterns are established during the stages of embryogensis[138-140]. Thus, across both plants and animals, DNA methylation is often considered as the hallmark of gene repression and silencing. However DNA methylation in invertebrate genomes has been controversial because the two post popular model organisms, *Drosophila melanogaster* and *Caenorhabditis elegans*, are mostly methylation free[134]. Comparative studies of other invertebrates from a diverse set of phyla showed DNA methylation patterns being present in their genomes[141] but the basal metazoans have been overlooked. Here, we present the first evidence for DNA methylation in ctenophores.

DNA methylation is the biochemical modification in which a methyl group is added to position 5 on the pyrimidine ring of the cytosine on the DNA molecule, creating the 5th base or 5- methyl cytosine (5-mC)[142]. The primary DNA sequence does not change by this addition, but the covalent modification of the DNA by methylation can have a profound impact on gene expression. The first mark of DNA methylation is the presence of the enzyme DNA methyltransferase (DNMT) that catalyzes the transfer of a methyl group to DNA. Consistent with the absence of detectable DNA methylation, yeasts (*S. cerevisiae*) and some ecdysozoans (*D. melanogaster*, and *C. elegans)* have no DNMT. In contrast, *Pleurobrachia bachei* has one DNMT-like enzyme (PbDNMT) that shares a high identity (52%) to the human DNMT1. However, phylogenetically the PbDNMT is recovered as the most basal to all metazoan DNMTs (**Extended Data Fig. 2**).

Structurally, the PbDNMT contains all the conserved domains of other homologous enzymes[143-147] including the N-terminus DMAP domain for secondary protein interactions, a

nuclear localization sequence (NLS), a zinc finger with a Cys-X-X-Cys domain that recognizes CpGs, and localizes the DNMT to DNA replication forks as well as two Bromo-adjacent homology domains used in protein-protein interactions. Importantly, the PbDNMT has a conserved DNA-methylase region that is subdivided into a catalytic core and a target recognition domain[148].

DNA methylation is often associated to the CpG context, therefore, facilitating the elimination of these CpG dinucleotides over evolutionary time[66]. It was shown that methylated cytosines tend to turn into thymines because of spontaneous deamination and a recovery enzyme, thymine-DNA glycosylase, or TDG, is not sufficient to effectively prevent the relatively rapid mutation of the dinucleotides leading to a relatively lower occurrence of CpGs in genomes with active DNA methylation[66]. For example, frequency of CpG dinucleotides in the human genome is ~1% - less than one-quarter of the expected frequency due to random chance[149,150]. The *P. bachei* genome contains 2.3 % CpG dinucleotides, which is also much lower than the expected random frequency and indicative to a genome that undergoes methylation compared to humans (**Extended Data Fig. 2a**, for details see **Supplementary Table 6S**).

The low frequency of CpGs and DNMT strongly suggested the presence of enzymatic DNA methylation and its product 5mC in the *P. bachei* genome. To experimentally validate this prediction, we used ELIZA based colorimetric assays (see the **Supplementary Methods SM8**) and quantified global 5-mC methylation in the *P. bachei* genome[57]. Surprisingly, we detect both 5-mC and 5-hydroxymethyl cytosine (5-hmC, the 6th DNA base, **Extended Data Fig. 2d**). Thus, *Pleurobrachia* represents the first reported invertebrate genome with detectable quantities of 5-hmC (**Extended Data Fig. 2d**)[57]. Importantly, the *P. bachei* genome contains the TET-like gene too. TET family of enzymes catalyzes the oxidative modification of 5-methylcytosine (5-mC) to form 5-hmC. In humans there are three paralogous proteins (TET1, TET2, and TET3) and these proteins have previously been described as the human LCX (leukemia-associated protein with a CXXC domain)/TET1 (ten-eleven translocation-1) and further designated as TET/JBP[151].

RNA-seq profiling indicates differential expression for DNMT and TET-like genes during development and in adult *P. bachei* (**Extended Data Fig. 2c**). Both DNMT and TET-like genes are predominantly expressed during the cleavage starting from the 1st division. However, the TET-like gene is also highly expressed in the adult combs (**Extended Data Fig. 2c**). For sequences and expression data see **Supplementary Excel Table 31S.**

SD4 Phylogenomic Analyses

The phylogenetic position of Ctenophora is elusive with numerous conflicting hypotheses generated over the last few years (see main text and[1-6,276-282]). Five basal lineages (Ctenophora, Porifera, Placozoa, Cnidaria and Bilateria) are representative of all extant metazoan diversity (**Fig. 1**) and understanding their evolutionary history is paramount for deciphering the early evolution of organ systems[77,152]. Thus, we used the emerging genome-scale data to clarify the relationships among ctenophores and other animal clades.

To reconstruct basal metazoan phylogeny, we conducted two sets of phylogenomic analyses including new data from *Pleurobrachia bachei* and the sponges *Sycon* (Calcarea) and *Aphrocallistes* (Hexactinellida) as well as existing data (see **Supplementary Table 12S**). A third analysis, also including new data from *Pleurobrachia bachei*, examining the influence of highly divergent ctenophore sequences was also conducted. This analysis differed slightly in taxon sampling due to gene selection.

For the first set of analyses, Ctenophora was represented by two species of *Pleurobrachia bachei* and *Mnemiopsis leidyi*. For a subsequent analysis, sampling within Ctenophora was expanded to include ten additional taxa, each represented by a relatively deeply sequenced *Illumina* transcriptome (**Supplementary Table 13S**). Other minor changes were made to improve taxon sampling. Additionally, we modified our bioinformatic pipeline to employ a more accurate translation utility, include stricter screening of potential paralogs, and require more taxa to be sampled per gene (see **Supplementary Methods SM7**).

SD4.1 Analyses with Ctenophora represented by *Pleurobrachia bachei* and *Mnemiopsis leidyi*

Our bioinformatic pipeline produced a final data matrix of 170,871 amino acid positions across 586 genes with 44 taxa. On average, taxa in this matrix were sampled for 51.3% of genes and overall matrix completeness is 45.4%. This matrix was analyzed on the Auburn University Molette Lab SkyNet supercomputer using maximum likelihood in RAxML 7.2.7[39] under the WAG+CAT+F model which has been demonstrated to perform similarly to using the best-fitting amino acid substitution model for each gene in phylogenomic analyses but is less computationally intensive [31]. Nodal support was assessed using 100 bootstrap replicates. Fungi were initially used to root the resulting topology. Leaf stabilities and branch attachment frequencies were examined using Phyutility [40].

Because leaf stability scores for the placozoan *Trichoplax* and the nematode *Caenorhabditis* were very low (0.8485 and 0.7998, respectively; **Supplementary Table 12S**) and examination of branch attachment frequencies showed that these taxa moved throughout Metazoa among bootstrap replicates (data now shown), the data matrix was re-analyzed (using the same methods) excluding these unstable taxa. Notably, the leaf stability scores for the hexactinellid and calcareous sponge OTUs were also low in the original 44-taxon analysis. However, examination of branch attachment frequencies showed that placement of these taxa varied primarily within Porifera (data not shown) and thus they were not excluded.

In order to demonstrate that placement of Ctenophora as the basal-most metazoan lineage is not the result of long-branch attraction to a distant outgroup, another analysis (also using the same methods) was conducted on a matrix excluding the two unstable taxa as well as all Fungi, *Capsaspora*, *Amoebidium*, and *Sphaeroforma* leaving only the two choanoflagellates as the outgroup used to root the tree (**Extended Data Fig. 3c**).

After Ctenophora, Porifera was recovered sister to all remaining metazoans (bs = 100%) with Cnidaria sister to Bilateria (bs = 100%, **Extended Data Fig. 3**). Relationships among major

lineages of Porifera were poorly resolved but relationships within Cnidaria and Bilateria were generally strongly supported and consistent with previous studies[153,154]. However, the cephalochordate *Branchiostoma* was recovered as sister to Ambulacraria (Echinodermata + Hemichordata) rather than with the other chordates. This result is likely an artefact of limited taxon sampling within Deuterostomia[155].

Our placement of Ctenophora (**Extended Data Fig 3**) echoes results from Dunn et al. (2008)[38] and Hejnol et al. (2009)[156] and Ryan et al (2013)[55] but is at odds with other phylogenomic studies placing Ctenophora sister to Cnidaria[157] or between Porifera and all other animals[154]. Although basal placement of Ctenophora, as recovered here, is in contrast to traditional views of metazoan phylogeny, analysis of the genome of *Pleurobrachia bachei* provides additional support for this phylogenetic hypothesis (see main text and the supplement sections with analyses of gene families and gene loss). It should be noted that we attempted to analyze this dataset using Bayesian inference but it was not computationally feasible for such a large dataset.

Hypothesis testing was conducted using the SH-test[89] based on taxon sampling corresponding to **Extended Data Figure 3c**. Both the Eumetazoa hypothesis (sponges outside of a clade including all other metazoans; likelihood=-3,285,686.94, D(LH)=-415.83, SD=41.02) and the Coelenterata hypothesis (likelihood=-3,286,647.50, D(LH)=-1376.39, SD=81.66) were rejected.

SD4.2 Analyses with expanded Ctenophora sampling

Taxon sampling for the phylogenomic analysis with expanded ctenophore sampling is presented in **Supplementary Table 13).** In addition to the genome of *Pleurobrachia bachei*, our taxon sampling included new transcriptome data from the sponges *Sycon* sp. (Calcarea) and *Aphrocallistes vastu*s (Hexactinellida) as well as ten additional ctenophore species (*Euplokamis dunlapae*, *Coeloplana astericola, Vallicula multiformis, Pleurobrachia pileus, Pleurobrachia* sp., *Dryodora glandiformis, Beroe abyssicola, Mnemiopsis leidyi , Bolinopsis infundibulum* and *Mertensiidae* sp., Supplementary **Tables 10-13S**) broadly spanning the diversity of the group.

Briefly, we built on the most reduced taxon sampling scheme described above by excluding the non-choanoflagellate outgroups and the long-branch taxon *Caenorhabditis elegans* and adding relatively deeply sequenced Illumina transcriptomes for nine additional ctenophore species. A small number of publicly available datasets were also added, in some cases replacing smaller datasets sampled in our earlier analysis.

A modified version of the bioinformatic pipeline used for our earlier analysis produced a final data matrix of 22,772 amino acid positions across 114 genes with 60 taxa. On average, taxa in this matrix were sampled for 48.2% of genes and overall matrix completeness for this data matrix was 51.8%. When we integrated transcriptome data from ten additional ctenophore species and modified our bioinformatic pipeline to be stricter at excluding paralogs (**Supplementary Methods**), the relative placement of ctenophores and sponges were less well supported. Ctenophora was still placed as the basal-most animal lineage and Porifera was recovered monophyletic and sister to all other metazoans besides ctenophores. Weak support for

relationships within Porifera and among basal metazoan phyla in this analysis may be due to limited data available most sponge lineages.

Hypothesis testing was conducted using the SH-test[89] based on taxon sampling corresponding to **Expanded Data Figure 3d**. The Coelenterata hypothesis was rejected (likelihood=-502,447.11, D(LH)=-310.11, SD=39.35) but the Eumetazoa hypothesis was not (likelihood=-502,145.26, D(LH)=-8.26, SD=11.19).

Additionally, we obtained strong support for relationships among sampled ctenophore lineages (bs = 90-100%, main text). Importantly, the reconstructed phylogeny strongly suggests multiple independent losses of both the cydippid larval stage and tentacle apparatus across Ctenophora with benthic Platyctenida (or aberrant ctenophores) as the second basal-most branch in Ctenophora. It appears that these aberrant ctenophores are secondarily simplified benthic animals that lost their swimming/pelagic stages, and convergently developed bilaterial symmetry. Interestingly, *Mnemiopsis leidyi*, a popular model species for developmental biology is also recovered as a relatively derived ctenophore lineage. To our surprise, an undescribed species, recently discovered by C. Mills and currently unnamed (listed as Mertensiidae sp. in the current manuscript), is superficially similar to *Pleurobrachia* but is recovered in a clade with the lobate ctenophores *Bolinopsis* and *Mnemiopsis*. In contrast, *Euplokamis* - one of the fastest swimmers among ctenophores - was recovered as the most basal in our analysis further supporting the hypothesis that the classical cydippids are polyphyletic.

SD4.3 Taking into account protein divergence for phylogeny inference

Maximum likelihood analyses raised the possibility that the position of Ctenophora as sister group to all other metazoans could correspond to long-branch attraction due to the high divergence of ctenophore sequences (**Supplementary Table 33S**). However, Bayesian analyses failed to resolve the phylogenetic position of ctenophores when only conservative genes were employed (data not shown). These results imply the need for future research and development of novel approaches to examine and control for the effects of highly divergent sequences.

SD4.4 Primary Absence of Many Animal-specific Genes in Ctenophores

We estimated the extent of global gene loss/gain across major animal clades (**Supplementary Table 14S**) using two alternative phylogenetic hypotheses: (i) Ctenophora as the earliest lineage sister to all Metazoa (**Fig. 1)** versus (ii) Sponges as the first branching animal lineage. Consistent with the recovered phylogeny (**Figs. 1, Extended Data Fig. 3**), placement of Ctenophore at the base of Metazoa provides the most parsimonious explanation of the data. Indeed, the desmosponge *Amphimedon queenslandica* shares more genes with bilaterians (11,674) and the sequenced sponge's genome[11] contains more evolutionarily conserved TF families including bZIP, T-box, Smad and immune genes than ctenophores. We repeatedly observed smaller gene complements for pan-animal gene families (especially those controlling development and signaling) for *Pleurobrachia* when compared to *Amphimedon* (**Supplementary Table 15S**). There are even less shared genes between ctenophores and cnidarians[12,68] or bilaterians (**Fig 3a,b, and Supplementary Table 14bS**) whereas 17,147 *Pleurobrachia* genes, or ~78%, were aligned to another ctenophore (**Supplementary Table 13S**).

Notable gene families absent in *Pleurobrachia* and ten other ctenophores are: (a) HOX genes involved in anterio-posterior patterning of body axes and present in all metazoans except ctenophores and sponges[80] (Supplementary **Table 17-18S**); and (b) canonical microRNA machinery (i.e. miRNA processing enzymes *Drosha* and *Pasha*, Supplementary **Table 19S**). In contrast, canonical microRNAs were readily detected in sponges, cnidarians and bilaterians[158]. The *Pleurobrachia* genome and other ctenophore transcriptomes contains only two nuclear hormone receptor genes, as also found in sponges, forming a distinct subfamily within Metazoa (see below), whereas there is a remarkable expansion of nuclear receptors in both cnidarians and bilaterians with at least 31 subfamilies, contrasted with their absence in choanoflagellates and fungi – sister groups of all animals[159,160].

Pleurobrachia also lacks the major elements that initiate innate immunity in other metazoans such as pattern recognition receptors (Toll-like, Nod-like, RIG-like, Ig-TIR) and immune mediators, MyD88 and RHD TFs, that are present in bilaterians, cnidarians and, in divergent forms, in sponges [161,162] (**Supplementary Tables 15S** and **20S** and below). On the contrary, 13 members of the immune ETS transcription factor family appear to be the result of an independent expansion within the *Pleurobrachia* lineage since direct orthologs in bilaterians were not found. In addition, we identified two genes that encode for peptidoglycan recognition protein (PGRP) domains, known as key players in the immune systems of protostomes and deuterostomes, which may have been secondarily lost in cnidarians and sponges (Supplementary **Tables 20S**).

Key bilaterian myogenic/mesoderm-specification genes are absent both in the *Pleurobrachia* genome and transcriptomes of ten other ctenophores (e.g. myogenin, canonical MyoD, troponin, paramyosin, z-disk components homologs, twist, nodal, FGF, gli/glis, Lbx, NK2,NK3, Myf4, Myf5, noggin, tinman, ladybird, BMP/dpp, Shh/hh, endomesoderm- however, *Pleurobrachia* and *Mnemiopsis* have GATA- **Supplementary Table 35S**). Consistent with our phylogeny these data suggest that muscles[163] and, possibly, mesoderm have evolved independently in the Ctenophore lineage to control the hydroskeleton, body shape and food capture. In contrast, ciliated locomotion itself, being preserved in extant ctenophores, reaches the most extreme specialization with cilia in ctenes approaching several millimetres – 200 times the length of cilia in other animals.

SD5 Illustrated Examples of Unique Genetic Information Processing in *Pleurobrachia bachei*

We outline selected gene families identified in both the *Pleurobrachia* genome and transcriptome dataset. Here we obtained deep, ~2,000X, transcriptome coverage across virtually all developmental stages and adult tissues, it allowed us to validate the gene model predictions as well as acquire quantitative expression information for most of the genes discussed below. We believe that this strategy represents a conservative approach in the initial analysis of this unique genome.

SD5.1 The Developmental Toolkit of Ctenophores.

There are conserved signaling pathways and transcription factors that are considered important developmental genes documented in the *Pleurobrachia* genome. Developmental genes work within gene networks, so entire genetic pathways were analyzed as completely as possible. *P. bachei* contains conserved developmental pathways of *Wnts*, *TGFßs*, and *Notch/Delta*, but lacks *Hedgehog*, and *FGF*. In bilaterians, expression of these developmental pathways is critical for early embryonic axis formation and many of the signaling pathways have duplicated into large gene families, as documented in the cnidarian, *Nematostella vectensis*[68,164].

Pleurobrachia has a reduced set of just 4 *Wnt* genes when compared to cnidarians, which contain over a dozen *Wnts*[164,165]. Most of the members of the *Wnt* pathway have been shown to be present in *Mnemiopsis leidyi*, although they are lacking some of the *Wnt* antagonists, such as *Dikkopf, Cerebrus*, and *Wnt* inhibitory factor[166], which are also lacking in the *P. bachei* genome (**Supplementary Excel Table 31S**). Most *Wnts* in the *P. bachei* genome group with related *Wnt* genes in *M. leidyi*. The *Wnt* pathway is expressed during development in *P. bachei* (**Supplementary Excel Table 31S**), but its role in the specification of tissues has not yet been elucidated. However, at least one gene – WntX is apparently associated with integrative and neural-like functions as evidenced by its high expression level in the aboral organ and major conductive pathways of *Pleurobrachia bachei* (**Extended Data Fig.5e.)**).

Ctenophores also have most *TGF-β* superfamily pathway members, but key agonists (thallostatin and chordin) are not present in the *P. bachei* genome, nor are they found in the *M. leidyi* transcriptomes (RNA-seq data) or its genome[166]. It is interesting that ctenophores lack antagonists of both the *Wnt* and the *TGF-β* developmental pathways, a feature that has also been reported in sponges[11] and *Trichoplax*[69]. These results suggest that the basic signaling pathways were elaborated in a metazoan ancestor, and that increased complexity, including the evolution of ligand-receptor antagonists, evolved later, before the split of the cnidarians and the bilaterians.

Developmental transcription factors also suggest that ctenophores are a very basal animal group, as they contain most of the transcription factor types found in all metazoans, but the number of each class of transcription factor is reduced when compared to cnidarians and bilaterians (**Supplementary Table 17S**). For example, the *P. bachei* genome has 92 homeodomain transcription factors (**Supplementary Table 18S**), but lacks the *Hox* cluster found in most bilaterians that is important for anterior-posterior patterning.

Pleurobrachia bachei is a direct developer with fast cleavage divisions occurring about every 30 minutes after the 1-cell stage. Cleavage produces a group of macromeres with the micromeres on top. The epiboly of the micromeres down over the macromeres and invagination of the stomodeum results in the gastrula stage (**Fig 4a,** main text). Series of RNA-seq profiling experiments indicate much higher overall transcription of ctenophore lineage-specific genes in the *Pleurobrachia* development, in particular during cleavage from 4-cell through 32-cell stages (**Supplementary Excel Table 31S**). Functions of these transcripts are yet to be elucidated. Transcripts encoding RNA-binding proteins are also abundantly expressed early in development. For sequences used in these analyses and expression data see **Supplementary Excel Table 31S**.

SD5.2 Transcription Factors

Transcription factors (TFs) are master regulators of transcription and, among other metazoans, *Pleurobrachia bachei* has an overall reduced complement of TFs. On the other hand, the *Pleurobrachia* TFs repertoire appears to be quite diverse compared to unicellular eukaryotes (e.g. *Monosiga brevicollis* and *Capsaspora owczarzaki*[167,168]) with examples of gene loss as well as gene expansions (**Supplementary Table 17S**). For example, a survey of the *Pleurobrachia* genome determined that several classes of TFs are absent in *P. bachei* but present in all metazoans and *C. owczarzaki*[168]. These classes of potential gene loss in the ctenophore lineage includes the Rel homology domain (RHD) TF, which includes a nuclear factor activated T-cell (NFAT) and Rel/NF-kappaB, the Churchill domain TF, a novel zinc binding TF.

SD5.2.1 **Homeodomain TFs.** We have identified a complete repertoire of 92 homeobox (HB) genes in the genome of *Pleurobrachia* (**Supplementary Table 18S**) - a significantly larger number compared to a more derived ctenophore species, *Mnemiopsis leidyi* [80]. Nevertheless, the canonical Hox TFs are absent in both species. Using a Maximum Likelihood (ML) tree reconstruction approach, these HB gene complement could be robustly grouped under 9 out of 11 metazoan HB gene classes (ANTP, PRD, LIM, POU, SIX, TALE, HNF, CUT, CERS, ZF, PROS)[169,170]. Two of these HB gene classes, such as ZF and PROS, were known as bilaterian-specific and not seen outside this lineage. Thus, analysis of *Pleurobrachia* HB gene complement suggests that the last common ancestors of all metazoans already had differentiated into all major metazoan-specific HB gene classes. Based on the orthology assignment with other metazoan HB genes, we infer that the last ancestor to all metazoans had at least 35 unique homeobox genes, more than previously anticipated. This number (35) also represents almost double the number (17) of HB genes that were inferred to be present at the last common ancestor to sponges[171]. We have detected several novel HB gene family members (e.g. Ems, Nk7.1, Hmx, BarH_b , Ceh-19, Cdx, Hlx, Vax, Pax4/6, Pax2/5/8, Prop, UncX, Mix) whose absence in the genomes of *Amphimedon* and *Trichoplax* generated the notion that they evolve later in evolution (Cnidarian+Bilateria). However, representative members of CUT, HNF gene classes pushed back the origin of these HB gene classes to the common metazoan ancestor.

Hierarchical clustering of homeodomain containing factors expressed in *Pleurobrachia* shows that adult combs have the highest overall level of expression (probably reflecting observed morphological and functional complexity in the organization of comb plates) followed by 8-cell through 64-cell cleavage stages. This parallels the level of transcriptional activity of RNA binding proteins **Supplementary Excel Table 31S**, implying that *P. bachei* may be co-opting RNA binding proteins for more extensive regulatory activities.

SD5.2.2. **The basic Helix-Loop-Helix** (bHLH) TFs. After homeodomain TFs, the second largest group of TFs is bHLH, which are TFs with a motif containing two α-helices connected by a loop[172]. Each stage of the *Pleurobrachia* development appears to have its own cohort of bHLH TFs **Supplementary Table 17S**). Interestingly, the TF bHLH-23 was also found to be most highly-expressed gene in the adult combs among all TFs. The *Pleurobrachia* bHLH 23 shares the highest identity to hairy and enhancer of split 1(HES). In mammals, HES is highly expressed in thyroid, prostate, and colorectal carcinoma[173].

SD5.2.3. **Nuclear receptors (NR).** Nuclear receptors (NR) are a large superfamily of receptors comprised of two broad classes of proteins, identified according to their mechanism of action and subcellular/nuclear distribution. These classes are further subdivided into four different types and at least 31 different groups of receptors[174]. NRs protein domain organization include a conserved DNA-binding domain (DBD) and a moderately conserved ligand-binding domain (LBD)[174]. NRs are not a metazoan innovation; even though they are absent from the genomes of choanoflagellates, fungi, plants, and prokaryotes, they are present in *C. owczarzaki* [167]. Among basal metazoans, it was determined that the placozoan *Trichoplax adhaerens* has 4 NRs and the cnidarian *Nematostella vectensis*, has 17 NRs[175]. Also two NRs have been identified in the sponge *Amphimedon queenslandica,* AqNR1 and AqNR2[176], and two were recently cloned from *Mnemiopsis leidyi,* MlNR1 and MlNR2[177].

Two NRs were found in the sequenced *Pleurobrachia* genome with their homologs in *Pleurobrachia pileus* transcriptomes. Phylogenetic analysis indicates that both *Pleurobrachia* NRs cluster with the *M. leidyi* MlNR1and the sponge AqNR2 forming possibly a new unique clade of NRs for basal metazoans. The second NR cluster contains one HNF4 subfamily (NR2A) member and share greatest sequence similarity with HNF4s from other animals[177] including diverse bilaterians, cnidarians, *Trichoplax, Amphimedon* and *Mnemiopsis* MlNR2. For sequences used in these analyses and expression data see **Supplementary Excel Tables 31S**.

SD5.3. Innate Immunity Genes

The quickly evolving genes that play roles in immunity are often difficult to identify, even in closely related species. The strongest diversifying pressures are on receptors that interact directly with pathogens, while the downstream signaling components and transcription factors tend to be more conserved[178]. To identify immune genes in the *Pleurobrachia* genome, we used a combinatorial, low-stringency domain-based approach as described in[179], in which we searched predicted gene models, translated genome sequence and transcriptome data for conserved domains that typify immune genes in other species. Given the rapid sequence diversification, primary sequence similarity is often low. Thus, we relied on the more conserved domain architectures that are identified through key conserved residues and protein structures. Analysis of the divergent ctenophore genome reveals a highly reduced immune system that lacks many of the major elements that mediate immunity in other metazoans.

Reduced immune complement. The *Pleurobrachia* genome does not contain homologs of Toll-like receptors (TLRs), IL-1 receptors or Nod-like receptors (NLRs) that are present in most bilaterian species and have been identified also in cnidarians and, in divergent forms, in sponges[161,162] (**Supplementary Tables, 15S** and **20S**). Homologs of MyD88, the immediate downstream signaling mediator in all other metazoan phyla, were also absent. We identified nine TIR domain-containing proteins (Table A below), but none of them encoded transmembrane regions, Ig domains or leucine-rich repeats that characterize TLRs and IL-1 receptor-like proteins.

Table A: TIR domain-containing proteins in *P. bachei*

TIR	Gene	Domain structure
1	scaffold00653_11 [5108 - 9073]	TIR(2):Efh(3)

2	scaffold04430_5 [2486 - 3076]	TIR domain only
3	scaffold10808.1|size1940 #FGENESH: 1	TIR domain only
4	scaffold1557.1|size21219 #FGENESH: 2	TIR domain only
5	scaffold20074.1|size811 #FGENESH: 1	TIR domain only
6	scaffold2746.1|size12179 #FGENESH: 3	TIR:HEAT
7	scaffold491.1|size50099 #FGENESH: 7	TIR:TIR:Efh(2)
8	scaffold8040.1|size3358 #FGENESH: 1	TIR:TIR
9	scaffold8372.1|size3074 #FGENESH: 1	TIR:TIR

The *Pleurobrachia* genome contains 56 genes that encode a total of 137 scavenger receptor cysteine-rich (SRCR) domains (**Supplementary Table 20S**). This is consistent with the multiplicity of SRCRs observed in other basal metazoans; the poriferan *Amphimedon queenslandica*, also exhibits a marked expansion of this family. Notably, two predicted genes encode peptidoglycan recognition protein (PGRP) domains. PGRPs bind to bacterial peptidoglycan and play key roles in the immune systems of protostomes and deuterostomes, but have not yet been described in non-bilaterian metazoans. One of these proteins consists of a single PGRP domain that resembles those of bilaterians. The other consists of two PGRP domains, a complement control protein (CCP) domain, a Kringle domain, and a transmembrane domain, which appears to be a unique domain combination. Two proteins contain alpha-2-macroglobulin domains. Not surprisingly, cytokines are particularly difficult to identify in this divergent species. Three homologs of macrophage migration inhibitory factor (Mif) were identified. We have also identified homologs of this cytokine in cnidarian, placozoan, and poriferan genomes. This ancient cytokine is a potent effector of inflammation in mammals, although its function has not been well-characterized in other phyla. No homologs of interleukins, interferon or TNF were found. Finally, there is a small family of seven perforin-like MACPF molecules.

The most conserved aspects of the immune system are transcription factors that regulate expression of genes involved both in immune response and also in development of immune cells among other functions. In bilaterians, proteins that contain the Rel homology domain, most notably NF-kB and rel, are central regulators of the immune response. No RHD was identified in *P. bachei*, despite extensive searching. We were able to find RHD in the sponge and cnidarian genomes, but it is either missing or too divergent in sequence to be identified in the ctenophore. Notably NF-kB homologs are secondarily absent in the derived genome of *C. elegans*. There are four homologs of Interferon Regulatory Factor (IRF) family in the *P. bachei* genome which function downstream of recognition in several mammalian innate signaling systems. There are 13 members of the ETS transcription factor family within the *P. bachei* genome. These genes appear to be the result of an independent expansion within this lineage; direct orthologs of the bilaterian sequences were not found. The genome also contains a single ortholog of the GATA transcription factor, which plays important roles in hematopoiesis in vertebrates and insects.

Antiviral Pattern Recognition Receptor and it's downstream signaling pathway. Detecting and responding to foreign RNA in the cytoplasm is a major component of antiviral activity. RIG-like receptors (RLRs) are related proteins that identify viral RNAs and initiate downstream immune responses through direct interaction with signaling protein MAVS (IPS-1)[180,181]. Humans and mammals have three RLRs, which collectively recognize a wide variety of viral infections[182,183]. RLRs recognize viral RNAs and activate downstream signaling using a modular

domain architecture. The C-terminal RNA recognition domain (RD) binds viral RNA, inducing a protein conformational shift, which allows twin N-terminal caspase activation and recruitment domain (CARDs) to interact with the signal-transducing protein MAVS and ultimately activate cellular immune response[184,185].

In most metazoan genomes there are at least 1 RIG-Like molecule is present, including the upstream twin CARD domains[186]. These genes and domains were not found outside animals, and initially they were considered as a vertebrate-specific novelty. We found both RIG-like and IPS-1 in the genome of *Nematostella*. However, the sponge, *Amphimedon queenslandica* has only two RIG-like molecules with strong viral RNA binding affinities, suggesting different downstream signaling mechanisms between these two clades. Our survey of the *Pleurobrachia* genome for the CARD domain and the RLR's did not result in any positive hits, suggesting either the loss of RLRs and downstream CARD signaling or, more likely their primary absence in this basal metazoan genome. CARD containing protein CED4/APAF-1 plays central role in cell apoptosis and induce downstream NF-kB activation through CARD-CARD interaction[187]. Using search for a conserved NB-ARC domain in the APAF-1, we also failed to find it the *Pleurobrachia* genome further suggesting that this component of apoptosis is missing in this lineage.

SD5.4. Small RNA Complement and miRNAs

Introduction. Small RNAs are now recognized as versatile players in gene regulation[188,189] and antiviral defense mechanisms. Several classes of small RNAs have been discovered including small interfering RNAs (siRNAs), microRNAs (miRNAs), and Piwi-interacting RNAs (piRNAs); each of them having distinctive biogenic pathways, modes of action, molecular targets and functions[190-192].

miRNA-mediated gene silencing uses a slightly different complement of effector proteins than the siRNA pathway[193,194] or RNA interference (RNAi, see below). miRNAs are synthesized by RNA polymerase II in the nucleus as long (up to 1000 nt) transcripts, called pre-miRNAs, and characterized by imperfect hairpin structures[193,194]. The dsRNA-specific endonuclease, *Drosha*, and its dsRNA-binding partner, *Pasha*, in *Drosophila* (DGCR8 in humans), process pri-miRNA into pre-miRNAs[193,194]. These pre-miRNAs are transported to the cytoplasm via an Exportin-5 dependent mechanism[193,194], where Dicer works with a dsRNA-binding partner, Loqs in *Drosophila* (TRBP in humans) to process pre-miRNA into mature, single-stranded miRNA and load it into an RNA silencing complex similar to RISC[193,194].

piRNAs are probably the largest and most abundant group of small RNAs found in both vertebrates and invertebrates, but not plants. piRNAs are reduced to 24-32 nt in length after undergoing primary processing from their larger ssRNA precursor forms. These ssRNA precursors are transcribed from repetitive intergenic regions called piRNA clusters[188]. After primary piRNAs are produced, they are then secondarily processed (and amplified) through a ping-pong pathway, which requires Piwi proteins[188]; we also found 2 genes encoded Piwi protein in the *Pleurobrachia* genome. piRNAs then form RNA-protein complexes through interactions with Piwi proteins[190-192].

Multiple pathways for small RNA biogenesis and diverse mechanisms of gene regulation suggest extensive parallel evolutions of RNA signaling as well as its multiple recruiting in various cellular functions. Evolution of miRNAs appears to be very mosaic based on the fact that canonical miRNAs are absent in *Trichoplax adhaerens*, and the pre-miRNAs of Porifera, Cnidaria and Bilateria have distinct sizes. Additionally, there is little conservation of miRNAs between poriferans, cnidarians and bilaterians, except for one *Nematostella vectensis* miRNA [195].

RNA interference (RNAi) in *Pleurobrachia.* RNAi is a eukaryotic molecular system that serves two primary functions: 1) gene regulation and 2) protection against selfish elements such as viruses and transposable DNA. Two different subtypes of *Dicer* endonucleases (*Dicer-1* and *Dicer-2*) mediate these two important functions. While *Dicer-2* is primarily responsible for antiviral activity by forming the complex (RISC) together with R2D2 and AGO-2, Dicer-1 performs gene regulatory function by forming a complex with *AGO1* and *TRBP2*. A Dicer molecule ancestral to all animals duplicated into *Dicer-1* and *Dicer-2*[196], in the stem lineage leading to animals, which is then followed by a loss of one or other in a particular lineages. Loss of *Dicer-1* and *AGO-1* as well as *Pasha* in the genome of *Trichoplax* has been implicated in lack of miRNAs important for gene regulatory functions[196]. The same studies also showed that the all the dicer gene copies that are present in the genome of *Trichoplax* are of *Dicer-2* type, which is important for antiviral activity. In the same way, in the genome and transcriptome of *Pleurobrachia*, we have identified only *Dicer-2* (**Supplementary Table 19S**), suggesting that *Pleurobrachia* either lost or never developed canonical miRNA-based gene regulation. All three *Pleurobrachia* Argonaute genes are similar to AGO2 type (**Supplementary Table 19S**). Surprisingly, our *in situ* hybridization experiments revealed that both *Dicer-2* and *Ago2* are expressed in the Aboral Organ and Polar Fields (**Extended Data Figs.5a,b**) implying a unique molecular organization of components of neural and sensory system in *Pleurobrachia*.

Small RNAs in Ctenophores. To experimentally survey the small RNA complement in ctenophores, we constructed cDNA libraries from the <40 nt RNA fraction in *Pleurobrachia* as well as from two other species of ctenophores, *Beroe abyssicola* and *Bolinopsis infundibulum*. A total of 18 million reads were sequenced using an Illumina MPS platform (NCBI, SRA accession number Project: SRP005925). Recovered small RNAs had two different size distribution patterns, with the majority of the reads corresponding to 21 nt, and fewer to the larger size of 24-25 nt. Of these 18 million reads, 2.5 million reads from the 21 nt RNAs fraction can be mapped to 49,114 different transcripts of the overall RNA-seq assembly. This suggests that these small ~21 nt RNAs are endo-siRNAs and thus a large fraction of *P. bachei* genes can be under control of the siRNA pathway.

We next screened the *P. bachei* small RNA reads for presence of evolutionary deeply conserved miRNA sequences, such as *miR-1* and *miR-124*, and the most ancient members such as *miR-100, let-7, miR-9*[197]. However, we found no significant hits even after allowing mismatches outside the seed region. Specifically, to discover potential miRNA candidate genes we used the miR-Intess approach applied previously for annotation of miRNAs in vertebrates[198], flies[199] and nematodes[200,201]. Similar to other approaches for miRNA annotation[195], the central requirement for classification of a particular genomic locus as a miRNA gene is presence of a fold-back (hairpin) structure and location of the majority of the reads within the arms of the

hairpin consistent with the 2 nt 3' overhang *Drosha*/*Dicer* signature[176,201]. Although more than 1,000 genomic regions folded into hairpins, none of them showed the *Drosha*/*Dicer* signature. Instead, they showed distribution of the reads through entire hairpin regions, suggesting biogenesis different from the canonical miRNA pathway. Although we found *Dicer-2* and showed its specific expression in the aboral organ, polar fields and combs (**Extended Data Fig.5a**), neither *Drosha* nor *Pasha* genes were present in spite of the extensive searches of the *Pleurobrachia* genome and transcriptomes (**Supplementary Table 19S**).

Thus, absence of these key genes of the miRNA pathway (*Drosha* and *Pasha*), along with absence of the deeply conserved mature miRNAs in small RNA sequencing data, combined with our failure to identify with confidence any miRNA candidate genes based on computational analysis of small RNA, suggest that *Pleurobrachia bachei* lacks the canonical miRNA pathway[190,191]. However, considering remarkable developmental and functional tissue organization other riboregulators and their corresponding pathways[202] might exist and expanded in ctenophores. Interestingly, *P. bachei* has only one *Dicer* gene while there are multiple *Dicer* genes in other basal metazoans (*T. adhaerens* has 5 and *A. queenslandica* has 4 *Dicer* genes [203]). *P. bachei* contains 3 *Argonaute* (*ago*) genes similarly to *A. queenslandica* but *T. adhaerens* has only one *ago* gene. All sequences can be found in **Supplementary Excel Table 31S**.

SD5.5 RNA editing

RNA editing is targeted alterations in nucleotide content of RNA. As a result, transcriptional output becomes different from its original genomic DNA template. Thus, RNA editing in mRNAs effectively alters the amino acid sequence of the encoded protein, possibly producing a different property or function to the altered protein. For example, RNA editing of the potassium channel in polar octopuses support their temperature adaptation [204]. RNA editing includes such processes as nucleotide insertions, deletions and changes of selected nucleotides in a RNA molecule[205,206]. The term RNA editing was first introduced in 1986 when it was reported that four uridine nucleotides were inserted into the mitochondrial *cox2* gene in trypanosomes[207]. RNA editing is different from RNA maturation in which functional RNAs are produced from such processes as 5′ capping, splicing, and 3′ polyadenylation.

In metazoans, most RNA editing is performed by deamination of nucleotide bases. There are two identified types of deamination editing in metazoans. First is the C-U editing found in the apolipoprotein B gene (AID/APOBEC), which in humans is performed by a cytidine deaminase family of enzymes[208,209]. This edit in the apolipoprotein B gene in humans generated a UAA, a stop codon, from the CAA sequence and a resultant truncated protein[208,209]. In mammals, C-to-U RNA editing is thought to be relatively less common. The majority of known mammalian RNA editing changes are of the second type of deamination, A-to-I, catalyzed by adenosine deaminase acting on the RNA (ADAR) family of proteins[210]. All ADARs share a similar structural domain organization, with a variable number of double-stranded RNA-binding domains (dsRBDs) on their N-terminus and a highly conserved catalytic deaminase domain at their C-terminus [210]. Humans have five isoforms that are generated from three genes, *ADAR*1, *ADAR*2, and *ADAR*3[210,211]. The ADAR3 has not been shown to be functional[211]. The ADAR mediate Inosine (A-to-I) editing, and during translation, inosine is decoded by the translation

machinery as if it were a guanosine[211]. Examples where this codon change produces a change in the primary sequences are known for a few proteins, such as the human glutamate receptor (GluR)[212] and serotonin receptor subtype (5-HT_{2C}R)[213], as well as the *Octopus* potassium channel[204]. Interestingly, most of this type of RNA editing is found in nervous systems.

From sequence homology searches another member of the adenosine deaminases superfamily was identified, the adenosine deaminases acting on tRNA (ADATs). ADATs which only contain the deaminase domain are involved in A→I editing of tRNAs at or near the anticodon position and are conserved in eukaryotes from yeast to man[214]. ADATs have been hypothesized to be the evolutionary ancestors of ADARs, then through acquisition of dsRBDs, the two domains worked together on a macromolecular substrate. It was suggested that cytidine deaminases acting on mononucleotides (CDAs) are more likely the evolutionarily predecessors to ADATs and consequently to ADARs[214].

RNA editing in *Pleurobrachia*. *Pleurobrachia bachei* has at least nine members of the deaminase superfamily – more than in any metazoan sequenced so far: 4 genes belongs to ADAR type with dsRBDs, 4 genes are members of the ADAT family of enzymes, and one ADAR-like gene, which is difficult to classify due to its highly divergent organization. A phylogenetic analysis of these *P. bachei* ADARs and ADATs indicated that PbADAR1 shares high identity to *C. elegans* ADAR2 and it clusters with other ADARs. The conserved deaminase domain of ADARs has a catalytic zinc center in which 4 residues are critical for catalysis; the proton donor glutamic acid residue (E) and three zinc binding residues, two cysteines (C) and a histamine (H) residue (data not shown**)**. The other two *P. bachei* ADARs cluster together with a unique *C. elegans* ADAR1. Interestingly, the *C. elegans* ADAR1 has no reactive site glutamic acid residue, yet is functional[215]. Two *P. bachei* ADATs cluster with *Capsaspora* and *Salpingoeca* ADATs. Our analysis is also consistent with other phylogenetic analyses that place ADATs basal to the ADARs[216].

RNA-seq expression profiles of *P. bachei ADAR*s, *ADAT*s and *CDA*s, show unique expression patterns for each transcript (**Fig. 4c,** main text). The Pb*ADAR*1 has a re-occurring mode of expression across the developmental stages while the PbA*DAR*2 transcript is differentially expressed in adult tissue. Both Pb*CDA*1 and Pb*CDA*2 are exclusively expressed in developmental stages. These unique expression patterns indicate that RNA editing is widespread in *Pleurobrachia* tissues with a diversity of developmental and adult functions. For sequences used in these analyses and expression data see **Supplementary Excel Table 31S**.

Unfortunately, we have only begun to elucidate the different types and functions of RNA editing. Thus, *Pleurobrachia*, as a representative of basal metazoans, offers unique opportunities to understand the emergence of complex RNA-signaling. One of the new functions might be involvement of RNA editing in regulation of the RNAi pathways and generation of animal complexity. Since dsRNA-binding proteins lack sequence specificity, it was hypothesized that RNA editing mechanisms interact with the RNAi pathways by sheer competition and, therefore, reduce efficiency of RNAi pathways[217]. Also addition of an inosine to the RNA strand creates bubble-like structures and can decrease endonuclease activity - *Dicer* resistance[217]. We suggest, that the expansion of RNA editing in *P. bachei* may be used as a complementary process to

control the RNAi pathways in the apparent absence of canonical miRNAs. This may be the reason why the *Pleurobrachia* lineage contains more ADARs than any other metazoan.

SD5.6 RNA Binding Proteins (RBPs)

RNA binding proteins (RBPs) are very versatile regulatory proteins with domains that enable them to bind single-stranded RNAs (ssRNAs) or double-stranded RNAs (dsRNA). Contacts between the RNA and protein can be made with the RNA bases, the ribose sugar and phosphate groups[218]. RBPs can recognize the primary RNA sequence or alternatively recognize RNA structure at a three-dimensional level[219]. Cytoplasmic and nuclear RBPs can regulate post-transcriptional events as well as translation of RNA[218,219]. To some extent RBPs are involved with almost all biological processes.

Using BLAST and Pfam Domain searches, we surveyed the *Pleurobrachia* genome and identified 260 RBPs with their distinct representation across protein classes compared to other basal metazoans and choanoflagellates examined[76] (see **Supplementary Table 21S**). Expression level of many RBPs, especially in developmental stages is very high and frequently comparable to such abundant transcripts as β-tubulin. Hierarchical clustering of the RBPs (RNA-seq) is very complex and each developmental stage has a unique expression patterns for RBPs (**Supplementary Excel Table 31S**). The 4-cell through 32-cell developmental stages have the highest overall RBP expression of suggesting that RBPs may be functionally co-opted in a regulatory role similar to transcription factors (but using much simpler binding the motifs than transcription factors[220-222],[223]). Below, we present selected examples of RBP motifs and classes in *P. bachei*. For an extended list of all the RBPs and sequences see the **Supplementary Excel Table 31S**.

SD5.6.1 RNA recognition motif (RRM) in RBPs

The RNA recognition motif (RRM) is present in up to 1% of all proteins across metazoans and consists of 80–90 amino acids which form a four-stranded antiparallel β-sheet with two additional α-helices arranged in-between giving a barrel-like topology[218]. The key feature of the RRM is its ability to recognize specific ssRNAs and interact with RNA motifs at the surface of the β-sheet in the RBP[218,224].

A specific class of RRM is the ELAV/Hu proteins that often recognized as pan-neuronal markers across Metazoa[225-228]. In *P. bachei,* there are 3 ELAV/Hu genes while the sponge, *A. queenslandica*, has only one (consistent with absence of neurons), and 4 genes were found in humans. Phylogenetic analysis shows that all 3 *Pleurobrachia* ELAV-like proteins form a basally branching cluster compared to their other homologs in metazoans, although all of them shared the same conserved domain structure found in other organisms. However, and in contrast to bilaterians and cnidarians, our in situ hybridization experiments show no correlation between ELAV/Hu and neural elements in *Pleurobrachia. For,* example, there are undetectable levels of ELAV-like expression in both subepithelial and mesogleal neural nets, the aboral organ (the analog of the elementary brain of ctenophores), polar fields, mouth areas, etc. (**Extended Data Figs. 5c,d***),* but non-neural elements in combs have very high levels of ELAV3 expression. RNA-seq profiling revealed that certain developmental stages also have higher levels of ELAV-

like expression *(***Supplementary Excel Table 31S***)* further suggesting that molecular makeup of ctenophore neurons can be significantly different that those in bilaterians and cnidarians, reflecting a parallel evolution of neural elements across basal metazoan lineages[152,229].

SD5.6.2 *K-homology (KH) domain in RBPs*

The KH domain is one of the most conserved RNA binding domains. Not surprisingly, this is also one of the most highly expressed RBPs in *Pleurobrachia*. The P. bachei genome has 31 predicted KH domain-containing proteins, more than in any basal metazoan examined, and the number of KH domains is extremely variable across RBP[230]. The KH domain can recognize both ssRNA and ssDNA, and is involved in a wide range of processes[230,231]. The KH domain is ~ 70 amino acids long and contains the signature sequence (I/L/V)IGXXGXX(I/L/V) in the middle of the domain[231]. There are two types of KH domains; type 1 which is found predominantly in eukaryotes, and type 2 which is mostly found in prokaryotes. Pleurobrachia has one RBP that contains a prokaryote–like type 2 KH domain in its 40S ribosomal protein S3.

One of the best-known metazoan RBPs with two KH domains is the Fragile X Mental Retardation Protein (FMRP) involved in learning and memory and a number of neurological disorders[232,233]. The sponge, *A. queenslandica,* and choanoflagellate, M. *brevicollis both have FMRP*. However, P. bachei does not have this gene in its genome suggesting its loss together with the Bicaudal-C genes. It should be noted that we found a distant homolog of FMRP-like gene in *Mnemiopsis*.

P. bachei also has a large expansion of the Neuro-oncological ventral antigen 1 *(NOVA)* gene family[234,235], totaling 6 NOVA-like genes – more than in any metazoan sequenced. The NOVA gene is not present in A. *queenslandica* and *M. brevicollis*. Humans have 2 *NOVA* genes and most invertebrates have only one. Phylogenetic analysis shows that the PbNOVA1 clusters with all other species and the other *P. bachei* NOVAs form a separate, basally branched sister clade. Expression patterns of *P. bachei NOVA* genes are quite variable with *NOVA*3 being most highly expressed in the developmental stages where *NOVA*2 in mostly highly expressed in the adult tissue (**Supplementary Excel Table 31S**). Also, the consensus binding site of NOVA, $(UCAU)_3$ [222] was found in the *P. bachei GluR12* gene indicating NOVA's possible involvement in regulation of neuronal activity. Interestingly, a homolog of the neural proliferation/differentiation control-1 (NPDC1) protein, which has been proposed to play a role in neural cell proliferation and differentiation, is most highly expressed in the adult aboral organ *Pleurobrachia*.

SD5.6.3 Other RBPs

Pleurobrachia bachei has a full repertoire of RBP domains including the DEAD/DEAH box helicase, VASA, PUF domain, *Pumilio*, and *Nanos*. *Nanos* is the maternal effect factor that contains a highly conserved zinc finger domain. *Nanos* has been implicated in development of the anterior-posterior axis as well as in germ cell and stem cell maintenance and migration. We found a single copy of *Nanos* in *P. bachei* (similarly to sponges, but there are the two *Nanos* genes in cnidarians[236]). *Nanos* expression levels fluctuate drastically during embryogenesis indicating its role in ctenophore development.

P. bachei also contains unique vault-type cytoplasmic ribonucleoprotein complexes. These large cytoplasmic ribonucleoprotein complexes are made of a major vault protein (MVP), two minor vault proteins (VPARP and TEP1), and several small untranslated RNA molecules of which are all present in *P. bachei*[200]. These multi-subunit complexes are thought to be involved in drug resistance as well as cancer[237]. Its function in *P. bachei* is yet to be deciphered. The complex of genes that code for the proteins in the large vault cytoplasmic ribonucleoprotein complex are expressed in developmental stages as well as combs, with the combs having the highest expression levels (**Supplementary Excel Table 31S**).

SD5.7 Extracellular Matrix Components (Collagens & Integrins)

Evolution of multicellular animals required development of epithelial tissues that support internal homeostasis and function to control bidirectional transport of molecules from environment to the organism. Collagens are the universal trait of epithelial organization and all Metazoa. We characterized the collagen complement from the genome of *Pleurobrachia bachei* focusing on the collagen IV - a key component of the Basal Lamina (BL). The BL is a layer of tissue which connects outer epithelial layers to tissues below and a necessary feature of any 'true epithelia[238]. Until recently, collagen IV had been detected only in Cnidaria, but in no other basal metazoans. It has been shown that Cnidaria use collagen IV in development and healing[239], and possess a basal lamina-like membrane containing collagen IV[240]. A recent study suggests that Porifera also has molecular components necessary to make a primordial BL, including collagen type IV[241].

Our findings provide new evidence that collagen IV is conserved across the entire metazoan lineage but a type of the IV collagen complement is the most unrivaled among basal metazoan groups in its diversity and complexity. Specifically, we discovered that *P.bachei* has seven distinct type IV collagen genes, an expansion unseen in any organism sequenced to date[242]. These genes show a unique distribution across the genome: two pairs are found in an in-line pattern (*αA* and *αB, αD* and *αE*); two are found independent of other genes (*αF, αG*), while another two are aligned in a head-to-head fashion (*αB, αC*) [242]. This exceptional arrangement suggests both traditional and inverted gene duplication events in the ctenophore lineage. Ctenophore collagen intron/exon arrangements were also uniquely diverse, with a range of 14-40 introns, depending on the gene. Phylogenetic analysis revealed two distinct collagen groups, the α1-like and α2-like sub-families that are common in bilaterians. Yet, gene-to-gene comparisons reveal a more diverse type IV collagen complement than found in vertebrate chordates, including humans.

Furthermore, we found differential expression in developmental stages between the sub-family gene types, indicating unique physiological use between these two collagen groups (**Supplementary Excel Table 31S**). Our findings imply that the common ancestor to all Metazoa might have contained a much more developed collagen complement than was previously appreciated. At the same time, there may have been extensive parallel evolution of ancestral collagens with remarkable functional specification, which influenced the diversity of body plans in ctenophores and their large body sizes.

Integrins are heterodimeric transmembrane receptors made of an α and a β subunit and are involved in cell-cell communication as well as extracellular matrix (ECM) interactions[243]. Integrins are not metazoan specific and have been found in unicellular organisms such as *Capsaspora owczarzaki*[244]. *P. bachei* has two α-like and five β-like integin subunits. Integrin interactions encompass many different pathways including neural-type semaphorin signaling and axon guidance. Semaphorin proteins contain a specific region of about 500 amino acids called the sema domain, and PSI (plexins, semaphorins and integrins) domains and are further distinguished by distinct protein domains[245]. Integrins as well as plexins are receptors to semaphorins[246]. The *P. bachei* genome contains 6 semaphorins, 8 plexin-like receptors and 2 ephrin-like receptors, completing the extensive receptor complement for semaphorin signaling and axon guidance pathways.

All sequences can be found in the **Supplementary Excel Table 31S**.

SD5.8 Genomic Survey for Neuronal Transmitter Candidates

Chemical neurotransmission and trans-synaptic communication are hallmarks of neural organization but their origins are elusive. About 30% of the human postsynaptic density proteins are conserved across majority of eukaryotes[247,248] suggesting that complex processing of environmental signals in unicellular organisms was recruited by early nervous systems[249]. This hypothesis might be reflected in the fact than significant majority of predicted synaptic proteins in *Pleurobrachia* is highly expressed during early embryonic stages, without any differential neurons or synapses (**Supplementary Excel Table 31S**). The sequenced *Pleurobrachia* genome provides the first insights into evolution nervous systems, and, here, we will primarily discuss the neurotransmitter and synaptic molecular complement in ctenophores.

Classical neurotransmitters are the low molecular weight intercellular messengers that transmit signals from a neuron to a target cell within or beyond a synaptic cleft. These transmitters are usually synthesized from amino acids precursors in a small number of biosynthetic steps. The transmitter is accumulated in synaptic vesicles and subsequently released by Calcium-dependent exocytosis mechanisms. Surprisingly, the *Pleurobrachia* genome and sequenced transcriptomes from other ctenophore species lacks all but one (glutamate/GABA) of the enzymes to synthesize any of the conventional low molecular weight neurotransmitters (**SupplementaryTable 22S**).

For example, recognized enzymes for synthesis of catecholamines are absent in the genome of *P. bachei*, including tyrosine hydroxylase, L-aromatic amino acid decarboxylase/dopa decarboxylase, dopamine beta-hydroxylase and phenylethanolamine N-methyltransferase. We did not detect catechol-O-methyltransferase, the terminal catecholamines inactivation enzyme. No biosynthetic enzymes were found for synthesis of tryptamine derivatives, including serotonin and melatonin, such as tryptophan hydroxylase. However, there is a phenylalanine hydroxylase related gene that phylogenetically clustered with phenylalanine hydroxylases but not tryptophan hydroxylase (**Supplementary Table 26S**). We found neither the histamine biosynthetic enzyme, histidine decarboxylase nor the degradation enzyme histamine N-methyltransferase. There is no molecular evidence for canonical acetylcholine synthesis and signaling due to absence of choline acetyltransferase. *Pleurobrachia* carnitine O-acetyltransferase-like enzyme is phylogenetically

clustered with carnitine O-acetyltransferase, but not with choline acetyltransferase (**Supplementary Table 27S**). We did find a mitochondrial type of monoamine oxidase B that shares highest identity with the bacterial form, and catalyzes the oxidative deamination of multiple biogenic and xenobiotic amines, therefore it might not be considered as a pure neurotransmitter related gene.

Pleurobrachia has no canonical Nitric Oxide Synthases[250], the synthetic enzyme for gaseous free radical messenger - nitric oxide (NO). However, we identified, the heme oxygenase, the enzyme, which catalyzes the degradation of heme to bilirubin and carbon monoxide (CO) but the bacterial-type carbon monoxide dehydrogenase was not detected. The enzyme cystathionine beta-synthase, which catalyzes the condensation of cysteine with homocysteine to form cystathionine and hydrogen sulfide (H_2S) is present suggesting that H_2S might be endogenously produced in *Pleurobrachia*.

Importantly, our metabolomic/microchemical assays (**Supplementary Tables 22S-25S**) also failed to detect most of classical low molecular weight transmitters (save glutamate/GABA) confirming the predictions of the current genomic survey. We conclude that there is no cholinergic, catecholaminergic, histaminergic or serotonergic signaling in *P. bachei* because of the absence of biosynthetic enzymatic pathways. Furthermore, none of the neurotransmitters from these pathways (such as serotonin, dopamine, noradrenalin, adrenalin, octopamine, histamine and acetylcholine) were detected by highly sensitive methods of capillary electrophoresis (**Supplementary Methods**).

SD5.9 Ligand-gated Ion channels

Ligand-gated ion channels (LGICs) or ionotropic receptors are transmembrane ion channels that are opened or closed in response to the binding of a chemical messenger such as a transmitter or signaling peptide or even metabolite and protons. There are four major classes of such ionotropic receptors: cys-loop receptors, ATP-gated channels, ionotropic glutamate receptors and acid-sensing ion channels (ASICs). The genome and transcriptomes of *Pleurobrachia bachei* do not have cys-loop receptors however there is a considerable diversity within other classes of receptors, significantly greater than previously described for any other basal metazoan (**Supplementary Tables 26S** and **27S**).

SD5.9.1 Ionotropic Glutamate Receptors

The amino acid, glutamate, is one of the most widespread intercellular messengers, and the most abundant molecule in nervous tissues. Glutamate receptors are divided into two major categories: ionotropic and metabotropic. The metabotropic glutamate receptors (mGluR) are G-protein coupled, and do not have ion channel domain – typically they mediate slower synaptic responses compare to ligand-gated receptors. On the other hand, the ionotropic glutamate receptors (iGluR) are cation channels mediating fast postsynaptic responses. In vertebrates, the iGluRs are further subdivided into classes based on the original names of their agonists. There are three classes of iGluRs, AMPA (α-amino-3-hydroxyl-5-methyl-4-isoxazole-propionate), KA

(Kainate) and NMDA (N-Methyl-D-Aspartate or NR) receptors. However this vertebrate classification does not correlate to the observed diversity of invertebrate iGluRs (see **Supplementary Table 27S** for the number and type of iGluR across phyla).

Remarkably, *Pleurobrachia bachei* has 14 iGluRs, more than all other metazoans studied to date. Genomic organization of these iGluR receptors is also more diverse than observed in other organisms: with exon numbers ranging from 2-20, there is no exon conservation compared to other metazoan iGluRs including KAs or NRs. We cloned all 14 *Pleurobrachia* iGluRs (GenBank accession numbers: JN202327-JN202339)[251,252]. All PbiGluRs have a conserved domain organization: they are composed of a signal peptide, an amino-terminal domain (NTD), a S1-S2 ligand-binding domain, three transmembrane domains, a reentrant pore loop and a carboxy-terminal cytoplasmic tail [253].

Using, *in situ* hybridization, we showed that most of PbiGluRs have highly differential and cell-specific expression pattern in the aboral organ, combs, polar fields (**Extended Data Fig. 7b**), with one of the highest expression levels in tentacles (**Supplementary Excel Table 31S**).

Phylogenetically, the *P. bachei* iGluRs with similar exon numbers cluster together, and it appears that this expansion of iGluRs in ctenophores is a result of lineage-specific duplication from a few ancestral prototypes[251,252].

Origins of NMDA receptors (NRs) are unknown but our analyses recovers ctenophore NMDA receptors outside of a clade consisting of Bilateria + Cnidaria (**Extended Data Fig. 7a**). It has also been shown that vertebrate NMDA receptors display a specific elongation of the intracellular C-terminus of the NR2 subunit[249,254]. Interestingly, we noted a similar type of an intracellular C-terminus extension for some *P. bache*i iGluRs. In particular, the Pb5 and Pb9 iGluRs have an especially long C-terminus that contains a CaMKII phosphorylation site and other post modification motifs and binding sites, as observed in the vertebrate NR2[249]. It was suggested that these unusual properties of the C-terminus of NR2 have evolved as a result of the increase in complexity of the postsynaptic signaling complex known to be coupled to NMDA receptor – the feature previously associated to the vertebrate lineage only[255]. Thus, the unusually long C-terminus properties of some *Pleurobrachia* iGluRs may also reflect the independent development of similar molecular complexity in the ctenophore postsynaptic organization conceptually comparable to vertebrate signaling.

SD5.9.2 Molecular Diversity of Glutamate Pathways in *Pleurobrachia*

A discovered diversity of iGluRs in both *Pleurobrachia bachei* and related ctenophore species correlates with a remarkable expansion of protein families associated with glutamate metabolism and processing. First, we noted a significant expansion in the complement of putative glutamate transporters (Glu transporters) represented by the Sialin superfamily. For example, the *Pleurobrachia* genome contains eight predicted Glu transporters compared to four similar transporters found in humans. Second, in *P. bachei* we found at least seven glutaminases, enzymes that convert glutamine to glutamate, whereas most of metazoans, including humans, have only two glutaminase genes.

These data combined suggest that observed diversification of glutamate processing is coupled with an enormous functional complexity of glutamate signaling in ctenophores, and most notably with the suggested role of L-glutamate as intercellular transmitter in Ctenophores. Interestingly, expression levels of genes associated with glutamate processing was substantially increased on the 3rd day of *Pleurobrachia* development when the first neurons appeared (**Fig. 4d**). Thus, the neurogenesis and the fate specification of glutamergic neural signaling might be inherently coupled with an orchestrated expression of all molecular components supporting the synthesis of L-glutamate (glutaminases), its inactivation (transporters/sialins) and targeted glutamate sensing by postsynaptic cells via iGluRs to support neuromuscular transmission for example.

SD5.9.3 Acid-Sensing Ion Channels (ASICs)

Compared to other metazoan genomes, there is a large expansion of acid-sensing ion channels (ASICs), with 29 detected in the *Pleurobrachia bachei* genome. For example, humans have only nine ASICs, and there are three ASICs in *Trichoplax* (**Extended Data Fig. 9**). These channels are part of a subfamily called ENaC/Deg. Pharmacologically, ASIC-like channels are shown to be targets of protons and a variety of peptides as well as other small molecules including amiloride and toxins[256-258]. We hypothesize that, in addition to L-glutamate (see main text and SD5.9.2), observed diversity of ASIC-like channels in *Pleurobrachia* was co-evolved with development of ctenophore-specific intercellular signaling with protons and novel secreted peptides as candidate transmitters (see below in the section **SD5.11**).

It is interestingly that the majority of ASICs start to be expressed by day 3 of development (**Supplementary Excel Table 31S**), which is correlated with morphological differentiation of recognized neurons (**Fig. 4d**). Although some of ASICs could act as sensors of pH or small ligands early in the development, many subtypes of the ASCIs are abundant in the adult tissues with the tentacles having the highest level of expression (RNA-seq). Functional properties of these channels would be important both to elucidate the unprecedented diversity of this interesting class of receptors in the ctenophore lineage and the evolution of transmitter signaling.

SD5.9.4 ATP-gated Channels in *Pleurobrachia*

Adenosine triphosphate (ATP) is an ancient signal molecule involved in both intracellular and extracellular communication. P_2X receptors are cation-permeable ion channels that open in response to the binding of extracellular ATP and belong to a larger family of the purinergic receptors. Initially, P_2X receptors were considered as vertebrate-specific but consequently they were found in the flatworm *Schistosoma mansoni*[259], the tardigrade, *Hypsibius dujardini*[260] and then the sea slug, *Aplysia californica*[60]. Finally, P_2X receptors were discovered in two very different eukaryotes (the amoeba *Dictyostelium discoideum* [261] and the green algae *Ostreococcus tauri* [262]) suggesting their early pre-metazoan origins.

We surveyed basal metazoan genomes to clarify relationships between different P2X receptor classes. Both cnidarians (*Nematostella vectensis* and *Hydra magnipapillata*) and sponges (*Amphimedon queenslandica*) contain two P_2X receptors each. However, we found only one P2X receptor in the ctenophore *P. bachei*. Our phylogenetic analysis recovers two distinct interior nodes, possibly inferring two ancestor molecular lineages of P_2X receptors. Representatives of the first lineage of P2X receptors are found in the green alga, amoebozoans and, among animals, only in Cnidaria and Ctenophora. The second group is composed from most of metazoan P2X receptors and only one receptor subtype in the choanoflagellate *Monosiga*. This bipartite phylogenomic clustering is reflected in two distinct patterns of the exon organization among P2X receptors further supporting the hypothesis of early divergence for two P_2X receptor lineages.

SD5.10 G-protein Coupled (GPCRs) and Tyrosine Kinase Receptors (RTKs) in *Pleurobrachia*

GPCRs comprise a large family of transmembrane receptors that sense a broad diversity of molecules[263]. They have seven transmembrane domains and are linked to a heterotrimeric G protein[263]. Signal transduction is initiated by GPCR binding a ligand causing a conformational change in the receptor that then activates the G protein. *P. bachei* has a complex repertoire of GPCRs, with little similarity to other species. Based on Pfam domain searches we identified 238 GPCRs in the *P. bachei* genome compared to 142 in *A. queenslandica,* and close to 800 in humans[11,263,264]. Many identified GPCRs might be candidates for putative neurotransmitter metabotropic receptors (**Supplementary Table 26S**) and their expression indeed correlate with the morphological appearance of first neurons at the 3rd day of development. However, due to low level of sequence conservation of GPCRs among basal metazoan phyla, their ligand specificities will need to be determined experimentally.

Receptor tyrosine kinases (RTKs) belong to a large superfamily of protein kinases. These RTKs contain a transmembrane receptor region with a tyrosine kinase domain inside the cytoplasm[265]. We identified 239 protein kinases in *Pleurobrachi*a of which 40 are RTKs. *A. queenslandica* has one of the largest groups of RTKs with more than 150, while humans have 58[11,266]. About two hundred recognized GPCRs and RTKs represent a very complex system of sensory transduction in *Pleurobrachia* possibly associated to chemoreception, development and neural signaling. Although, we do not know the ligands for these orphan receptors their molecular organization is similar to many neuropeptide receptors in other animals suggesting that endogenous secretory peptides can be produced as transmitters in ctenophores.

SD5.11 Putative Peptidergic Signaling in *Pleurobrachia*

Peptidergic signaling involves use of small peptides as transmitters or hormones. Targets of these small peptides are usually G-protein coupled receptors (GPCRs) and ASIC-like receptors - both of them are present in *Pleurobrachia*. Signaling (neuro)peptides in animals are produced from large precursors or prohormones with a set of processing enzymes such as serine protease proprotein convertase 2 (PC2), proprotein convertase 1 (PC1/3), and the subtilisin-like proprotein convertase, furin.

In the *Pleurobrachia* genome and transcriptomes we do find two of these proteolytic enzymes: (i) PC1/3, which converts unprocessed peptides to intermediate substrates and (ii) furin, which generates the biologically active products. However, PC2 was not detected. Importantly, we also found the peptidylglycine alpha-amidating monooxygenase (*PAM*), which catalyzes precursor peptides to active alpha-amidated products. Interestingly, the adult combs and tentacles have the highest level of expression (RNA-seq data) for PC1/3 and furin suggesting that majority of secretory peptides are processed in these organs, also known as true ctenophore innovations. In contrast, PAM is predominantly expressed in development. Although, all three processing enzymes are expressed in re-occurring patterns, 16-cell and 1-day stage embryos had their highest expression levels (**Extended Data Fig. 8c**).

Identification of actual prohormones is a very challenging task, due to the enormous molecular diversity of these molecules and a limited numbers of conserved features such as presence of signal peptide motif[27,28], dibasic and other cleavage sites[267-270], repeat sequences, etc. Because almost all pre-secreted peptides are localized to the endoplasmic reticulum via the presence of an N-terminal signal peptide consisting of 5-20 hydrophobic residues, this heuristic forms the strongest marker of secreted products[27]. Complicating the prohormone discovery process is the presence of products containing one or more transmembrane domain that are directed to the ER but not secreted, and instead are embedded in the membrane[28]. Hidden Markov model, TMHMM[29], is employed to predict transmembrane topology and Phobius[30] uses complementary hidden Markov model, trained separately for transmembrane domains and signal peptides, to call a consensus. Although these prediction tools have been tested in a wide range of model organisms, plants, and bacteria, their validation has not been considered for basal metazoan groups such as ctenophores.

In order to achieve an initial prediction in the absence of prior studies for ctenophores, we developed three non-overlapping sets of results using different cuttoff criteria (**Supplementary Table 28S**). For an initial genomic survey of possible neuropeptide candidate, we developed a custom computational pipeline outlined in **Extended Data Fig. 8a** that employs several software packages to screen for molecular features in small peptide processing such as the presence of signal peptide motifs indicating secretion, lack of other subcellular localization motifs, absence of transmembrane domains, presence of at least one or multiple basic cleavage sites, cysteine patters[20 271 272] (see also **Supplementary Method section S4.2.3.7**).

Consequently, the workflow was designed to simultaneously predict secretory signaling peptides across other ctenophore species we sequence here (**Supplementary Tables 3S** and **10S**) – thus presence of homologs for new putative precursors in related ctenophores was also used as an independent confirmation of a correct prediction for novel ctenophore lineage-specific genes. As a result, we identified 499 putative secretory products (**Supplementary Table 32S)**, all of them were manually examined and cross validated in terms of their cell-/tissue specific expression (RNA-seq), relative abundance, predicted processing sites, and presence of related homologs or motifs in other species. Not surprising, species from the same ctenophore genus contained the most predicted *P. bachei* secretory peptides. Transcriptomes of benthic ctenophores, *Coeloplana astericola* and *Vallicula multiformis*, as more derived species, showed a lower degree of identity with many *P. bachei* secretory pepetides.

Thus, under the most stringent conditions, we identified 72 prohormones in *P. bachei*, and shown that most of them are predominantly expressed in adult tissues, although about 20 of them are differentially expressed early in development (**Extended Data Fig. 8c**). Tentacles display the highest expression levels, but each of the examined peptide precursors has its own distinctive expression profile. These findings were validated by *in situ* hybridization for selected precursor products revealing remarkable cell-specificity in their expression patterns (**Extended Data Fig. 8b**). We did not find any single ortholog of predicted ctenophore secretory prohormones or peptides to any identified neuropeptide/hormones in other metazoans.

SD5.12 Gap Junctions and Electrical Synapses

Gap junction proteins[273-275] are metazoan innovations. They are broadly divided into two groups: connexins, which are apparently vertebrate-lineage-specific genes; and pannexins, which present in both vertebrates and invertebrates. Before its discovery in humans[275], pannexins in invertebrates were initially named as innexins - i.e. the invertebrate counterpart to gap junctions. Innexin/pannexin types of gap junction proteins are each comprised of four transmembrane domains in which six individual innexins assemble to form a channel, an "innexon", in the plasma membrane[276]. Two opposing innexons on different cells form a functional gap junction. All gap junctions mediate fast electrical coupling between cells via electrical synapses. Interestingly, *Amphimedon*, *Trichoplax*, and *Nematostella* all lack recognized gap junction proteins. In contrast, the genome of *Pleurobrachia bachei* has 12 predicted pannexin/innexin-like proteins, which is a very large expansion of gap junction proteins for any metazoan group (**Supplementary Excel Table 31S**). For example, humans have only 3 pannexins.

Innexins are one of the most highly expressed transcripts in the adult aboral organ of *P. bachei* (**Fig. 5d**) but they are also detected in combs, conductive tracts (not shown) and in neuronal like subepithelial cells. In ctenophores gap junctions were previously identified by electron microscopy in the ciliated grooves which run from the apical organ to the first comb plate of each comb row, as well as through the endoderm of the meridional canals[277] suggesting that the major form of synaptic activity in the aboral organ is electrical. Gap junctions also allow the direct exchange of small molecules among neighboring cells and sometimes form channel aggregates in the plasma membrane that might explain expression of selected types of innexins during *Pleurobrachia* early development (**Fig. 5**).

In many animals, electrical synapses recruit fast acting voltage-gated channels that communicate along polarized cells through action potentials generated by voltage-dependent sodium (Na_v) channels. *P. bachei* contains two Na_v channels which share high identity to the two *Mnemiopsis* Na_v channels[278]. Porifera has no Na_v channels.

SD5.13 Ion channels

The common feature of channel proteins is the ability to form an ion-permeable pore through a biological membrane. Ion channels can be broadly classified by the nature of their gating, the number of pores (gates), and ion permeability or selectivity[279]. The gating, or opening

and closing of ion channels, can be further classified as voltage-gated, ligand-gated, cyclic nucleotide-gated, light-gated, temperature-gated or mechanosensory or leakage/water channels[279]. The fast acting, voltage-gated channels are primarily involved in excitability and long-distance communications (e.g. action potentials) among cells.

The ctenophore *Pleurobrachia bachei* has 112 recognized ion channels compared to 244 channels in humans, 191 in *Nematostella vectensis*, 50 in *Amphimedon queenslandica*, 48 in *Trichoplax adhaerens*, and only 22 in *Monosiga brevicollis* (**Extended Data Fig. 9**). Some classes of ion channels are apparently absent in the *P. bachei* genome. For example, there are no sodium leak channels (NALCN), no ATP-sensitive, or G protein-coupled inwardly-rectifying potassium channels, whereas sponges have two G protein-coupled inwardly-rectifying potassium channels[280]. **Extended Data Fig. 9** summarizes a hierarchical clustering of expression among all identified ion channels in both developmental stages and adult tissues. Although, we observed a unique expression pattern in every stage and tissue, adult tissues show significantly higher diversity and higher overall level of expression for most of ion channel types. In particular, mobile tentacles displayed the highest expression in Ca_v and Na_v.

P. bachei has 33 voltage-gated channels compared with only 5 in the sponge *A. queenslandica*, of which none are Na_v. The neuronal (N) type voltage-gated Calcium (Ca_v) and 2 Na_v channels share significant identity to the *Mnemiopsis leidyi* Ca_v and Na_v channels[278]. The unusual pore sequence D/E/E/A of both the *P. bachei* and *M. leidyi* Na_v channels was shown to be responsible for sodium selectivity[281]. Interestingly, we found very few accessory voltage-gated ion channel subunits in the genome of *P. bachei.*

Pleurobrachia has 3 cation channels also known as Catsper channels. In mammals, these two-pore channels are crucial for sperm hyperactivation and male fertility[282,283]. As in humans, there are 3 additional two-pore calcium channels (TPC) in the *Pleurobrachia* genome[284]. In mammals, TPCs have been shown to mediate nicotinic acid adenine dinucleotide phosphate (NAADP) dependent calcium release from acidic organelles[285,286]. High expression of the CapSper and Ca2P channels, particularly the Ca2P1 channels was found early in *Pleurobrachia* development (**Supplementary Excel Table 31S**).

The *Pleurobrachia* genome encodes a large diversity of voltage-gated potassium (K_v) channels, specifically $K_v\alpha$1-like (Shaker-like) channels totaling 25, compared to only 7 in humans and none in the sponge *A. queenslandica*. There are 2 $K_v\alpha$11-like channels, compared to 4 in humans and none in *A. queenslandica*. Interestingly, the development stages express much higher levels for $K_v\alpha$1.1a and $K_v\alpha$1.1c vs adults, while ttentacles express the most abundant numbers of $K_v\alpha$1.1s (**Supplementary Excel Table 31S**). In addition, we identified 5 tandem pore domains and 3 calcium-activated potassium channels.

Pleurobrachia has 9 transient receptor potential (Trp) channels that could be involved in mechanosensory activity. Trp channel expression is predominantly found in adult stages with only the Trp5 being detected in developmental stages. Surprisingly, 11 aquaporins or water channels were found in *Pleurobrachia*, which is far more than any other basal metazoan. Finally, we detected one cGMP-gated cation channel α-1-like, one ORAI calcium release-activated calcium modulator channel and 3 voltage gated chloride channels.

In conclusion, the *Pleurobrachia* ion channel repertoire is one of the most distinct among metazoans probably reflecting a combination of basal features and derived functional innovations associated with its predatory lifestyle and complex neuromuscular organization.

A highly reduced complement of animal-specific genes is a feature shared for the entire ctenophore lineage. Thus, ctenophores might have independently developed complex phenotypic plasticity and tissue organization, raising questions about the nature of existing ctenophore traits such as unique development, combs, tentacles and the apical organ.

SD 14. Parallel Evolution of Neural Organization in Ctenophores

Here we expand our analysis related to genomic bases of neuronal organization in Ctenophores and nervous system evolution.

Ctenophores are mostly pelagic predators with a remarkable behavioural repertoire including the most sophisticated ciliated locomotion in the animal kingdom (e.g. rapid coordinated reversal of different cilia bands) and prey capture with tentacle/colloblast apparatus and mesoderm-derived, muscles[77-79]. All effector systems are under control of the apical/aboral organ (an elementary brain with gravity sensors and a statolith consisting of about 100 lithocytes) and two distinct neural populations: the ectodermal hexagonal-type neural net and the more diffuse mesoglea network of neurons (**Fig. 5a, Extended Data Fig. 1**). Interneuronal and neuro-effector chemical synapses in all ctenophores have unique organization, never observed in other animals, forming a 'presynaptic triad'[78] with highly heterogeneous populations of vesicles suggesting the presence of multiple low molecular weight and peptide-type transmitters.

As most neurotransmitters and their synthesis pathways are highly conserved across bilaterians and cnidarians, we anticipated that ctenophores would share their neurotransmitter organization with other Eumetazoa together with majority of neurogenic and synapse-related genes. However, and in contrast to observations to all other animals with nervous systems, many genes controlled neuronal fate and patterning (e.g. Neurogenins, NeuroD, Achaete-scute, REST, and HOX/otx) are absent in the ctenophores we sampled. Orthologs of pre-and postsynaptic genes also have a reduced representation (**Supplementary Table 34S**) 'missing' components critical for synaptic function in other eumetazoans (i.e. organisms with nervous systems).

Importantly, our combined molecular, ultra-sensitive metabolomic, immunohistochemical and pharmacological data strongly suggest that ctenophores do not use acetylcholine, serotonin, dopamine, noradrenaline, adrenaline, octopamine, histamine or glycine as intercellular messengers (**Extended Data Figs. 6,7, Supplementary Data SD5.8**, **Tables 22-24S**). Lack of ionotropic receptors for these molecules in the *Pleurobrachia* genome and other ctenophore transcriptomes is consistent with this conclusion (Supplementary **Table 26aS**). The majority of synthetic genes for neurotransmitter pathways are also not present in sequenced unicellular eukaryotes (such as *Monosiga* and *Capsaspora* recognized as sister groups for animals) suggesting they are cnidarian/bilaterian innovations. *Pleurobrachia* also lacks NOS - a key synthetic enzyme involved in gaseous signalling mediated by nitric oxide[250]. In other words, ctenophores have the most dissimilar transmitter organization among all animals studied so far. But, what are the ctenophore transmitters?

We detected L-/D-glutamate and L-/D-aspartate in all ctenophore samples investigated (**Supplementary Tables 22-25S**). Physiological and pharmacological tests suggest that L-glutamate can act as a candidate neuromuscular transmitter in *Pleurobrachia* (**Fig. 5b**, **Extended Data Fig. 7**), able to induce rapid inward currents and rise intracellular Ca^{2+} in muscle cells (not shown) as well as cause muscle contractions at nanomolar concentrations (10^{-7} M). In contrast, all other classical neurotransmitters were ineffective even in concentrations up to 5 x 10^{-3} M while D-glutamate as well as L-/D-aspartate have significantly less affinity in these assays (**Fig. 5b**).

Although we did detect Gamma-Aminobutyric acid (GABA, a primary inhibitory neurotransmitter in the vertebrate brain, Supplementary **Tables 22-24S**), lack of pharmacological effects of GABA on *Pleurobrachia* behaviors and major motor systems such as cilia, muscle and colloblasts suggest that GABA acts as a passive by-product of glutamate metabolism by glutamate acid decarboxylase (GAD). Our surprising finding of GABA accumulation in muscle cells (but not in neurons as in other metazoans) implies that GABA may be a metabolic intermediate to inactivate the action of glutamate at the neuromuscular synapse. Interestingly, a product of GABA metabolisms itself can be a usable source of energy in ctenophore muscles. Indeed, GABA transaminase, also found in *Pleurobrachia*, is the enzyme that catalyzes the conversion of GABA back into succinic semialdehyde and glutamate following formation of succinic acid that enters the citric acid cycle – the universal aerobic bioenergetics pathway.

The hypothesis that glutamate is one of the signal molecules in ctenophores is supported by an unprecedented diversity of ionotropic glutamate receptors, iGluR, in the *Pleurobrachia* genome (**Extended Data Fig. 7a,b**) – far exceeding the number of genes encoding iGluRs in other basal metazoans[287]. iGluRs might have undergone a substantial adaptive radiation within the ctenophore lineage as evidenced by unique exon/intron organization for many subtypes and the topology of the tree forming a distinct branch within the entire family. Interestingly, in *Pleurobrachia* development neurons are formed two days after the initial muscle formation, and first neurogenesis events correlate with co-expression of all iGlu receptors in hatching larvae (**Fig. 4d**). All cloned iGluRs also show remarkable cell-type specific distribution with predominant expression in tentacles, followed by combs and the apical organ, revealing well-developed Glutamate-signalling in adults (**Extended Data Fig. 7**). In addition, the *Pleurobrachia* genome contains a greater assortment of enzymes involved in glutamate synthesis (8 glutaminases) and glutamate transporters (8 sialins) - more than any other metazoan investigated[288,289]. Their differential expression can reflect the functional diversification of glutamate synthesis and uptake.

The first nervous systems have been suggested to be primarily peptidergic in nature[152]. Although we did not find any previously identified neuropeptide homolog, the secretory peptide prohormone processing genes are present. Using a custom-designed computational pipeline, we predicted 72 novel peptide prohormones in *Pleurobrachia* and found more than 50 of their homologs in other sequenced ctenophores (**Extended Data Fig. 8**, Supplementary **Table 28S** and **32S**). These prohormone-derived peptides could have a variety of functions including cell to cell signalling, toxins or involved in innate immunity, or a combination. As many prohormones are rich in putative Cys-Cys bonds, this may suggest (in analogy to cone snail venoms) that some are toxins, an interesting area for follow-up studies. Several of these ctenophore-specific precursors are expressed in polarized cells around the mouth, in tentacles and polar fields,

suggesting a signalling role (**Extended Data Fig. 8b**). They can be natural ligands for more than 100 orphan neuropeptide-like G-protein coupled receptors[290], also identified in *Pleurobrachia*. The second notable example of neuropeptide receptor candidates is amiloride-sensitive sodium channels (ASIC), which are also known to be regulated by different classes of short peptides and protons[258,291,292]. The *Pleurobrachia* genome has 29 genes encoding ASICs - more than any organism sequenced so far (e.g. humans have only 9 ASICs), and expression of most of them is correlated with the morphological appearance of neurons in development. ASIC expression is most abundant in tentacles, combs and apical organs –structures highly enriched in neural elements and under complex synaptic control.

A generalized chemical synapse in ctenophores has a mosaic combination of ancestral and derived features yet with a reduced representation of orthologs of bilaterian/cnidarian pre- and postsynaptic genes. For example, *Pleurobrachia* and ten other ctenophores lack neuroligin (**Supplementary Table 34S**), but have a basal type of neurexins – a key component bringing together pre- and postsynaptic membranes in bilaterians[293] (yet lost in *Nematostella*). Surprisingly, predicted synaptic proteins are consistently expressed during early development in the absence of recognized neurons suggesting their additional functions as components of ubiquitous secretory and receptor machineries in eukaryotes.

In addition to chemical transmission, ctenophores evolved enormous diversity of electrical synapses (absent in *Nematostella,* sponges and placozoans) with 12 gap junction proteins (pannexins/innexins[273,275] but not chordate-specific connexins) encoded in the *Pleurobrachia* genome – all of them have their highest expression in the apical organ followed by tentacles and combs (**Fig. 5d**). The apical organ, combs and tentacles have a relatively large diversity of ion channels (**Extended Data Fig. 9b**), confirming complex regulation of excitability in these structures. We do not find any gap-junction orthologs in choanoflagellates or other basal eukaryotic groups suggesting that these are metazoan innovations with the profound expansion of this family in the ctenophore lineage. Yet the overall complement of voltage gated ion channels in ctenophores is reduced compared to other eumetazoans[278] (**Extended Data Fig. 9a**). *Pleurobrachia*, however, has voltage-gated sodium and many potassium channels that were apparently lost in sponges and a greater diversity of aquaporins (water channels[294]) then all other basal metazoans combined (**Extended Data Fig. 9**).

SD 15. Discussion: Did Neural Systems Evolve More Than Once?

The emerging new data from ctenophores allow us to revisit two scenario of neuronal evolution (**Extended Data Fig. 10**): (*i*) polygenesis or independent origins of neural systems in ctenophores vs cnidarian/bilaterian clade neurons[152,229,295-297], and (*ii*) monophyly or a single origin of the neural system with massive loss of majority neurotransmitters and some neurogenic molecular components in ctenophores. The corollary of this 'single-origin' scenario would be the catastrophic loss of the entire nervous system in both sponges and placozoans but the preservation of the original molecular makeup in bilaterian/cnidarian clade. We favour polygenesis hypothesis since many components of the molecular machinery controlling (1) neurogenesis, (2) neurotransmitter synthesis, (3) receptors pathways, (4) pre- and postsynaptic genes (including neuroligins and neurexins) are also absent in unicellular eukaryotes recognized

as sister groups of animals. (5) Given the current placement of ctenophore as the most basally branching animals, polygenesis of neurons seems the more plausible hypothesis for the origins of neuronal systems. Below, we will further clarify three controversial points following the discussion on the origin and early evolution of neural systems.

First, ctenophores lack components of canonical (neuro)transmitter systems – the feature well preserved in all eumetazoans including species with compact/'lossy' genomes (e.g. nematodes and ascidians) as well as all in parasitic animals investigated so far. From a number of genes encoding transmitter synthesis and degradation, only orthologs of genes distantly related to phenylananyne hydroxylase (PH) is shared between *Monosiga*, *Dictyostelium* and *Pleurobrachia*. Yet, as our capillary electrophoresis data suggest, these PH-related enzymes, if functional, do not produce any know catecholamines in ctenophores. Moreover, the neural transmitter systems are not only the presence of specific synthesis enzymes – the heterogeneity of secretory specificity of neurons is one of the most fundamental features of any nervous system. Transmitter phenotypes include very complex packing, uptake (transporters) and inactivation systems as well as multipart receptor machinery with several hundred of genes precisely co-expressed in a given neuronal cell type. Thus, the alternative possibility - the massive loss in ctenophores of virtually all receptor and transmitter synthesis pathways' genes is, in our opinion, a less parsimonious scenario. All ctenophores are predators with complex behaviours and a nearly complete replacement of multiple signalling pathways, as in single-origin hypothesis, seeps to be a more complex reconstruction.

Second, we also found that orthologous of bilaterian and cnidarian 'pan-neuronal' markers are not expressed in ctenophore neurons suggesting they perform different, possible ancestral roles. On the other hand, we identified two broad categories of other genes which are specifically expressed in ctenophore neurons. One and the largest the category includes genes that are either ctenophore innovations (such as ctenophore specific secretory peptides, Extended **Extended Data Fig. 8b**); or ctenophore lineage-specific isoforms (such as WntX; **Extended Data Fig. 5e**). The second category is a group of genes that are specifically expressed in the *Pleurobrachia* neurons but they are not specifically expressed in neurons in other eumetazoans (e.g. Argonaut or Dicers, **Extended Data Fig. 5a,b**). Combined, these data suggest that these genes are independently recruited to ctenophore neuronal machinery. An alternative explanation is that these genes or functions also were present in the common ancestor but were lost in all other metazoans; again this explanation seems more complex and less likely given conservation of neural components across animals.

Third, we do find a number of genes encoding pre- and post-synaptic proteins in the *Pleurobrachia* genome that are shared with Choanoflagellates and *Capsaspora* suggesting single-cell origin of the backbone of the canonical synaptic machinery revealed in proteomic studies on bilaterians[298]. However, our data indicate that these genes are not 'pure neuronal/synaptic' genes because even in ctenophores they are also expressed in non-neuronal tissues including secretory cells. Most interesting, the majority of canonical 'synapse–related' genes are highly expressed very early in development starting with the first cleavage when no neurons are present. This feature indicates that these "synaptic complex" genes are part of ubiquitous secretory and signalling complex that couples Ca-signalling with cell cytoskeleton organization as evidenced by their presence in unicellular eukaryotes. Thus, most of these

'synaptic complex' genes might not be considered alone as an evidence of early neuronal origin in the common metazoan ancestor but rather as the evidence for deep ancestral roots for such machinery in Opisthokonta/Holozoa. In addition, the synaptic gene complement in Ctenophores is significantly reduced compared to all other animals suggesting that ctenophore synapses are distinct in their molecular makeup as supported by electron microscopic investigations[78]. Combined, these data support our hypothesis about parallel evolution of synaptic organization in ctenophores which also reflects the development of their unique neurotransmitter organization**.**

Similarly, ion channels cannot be considered as specific neuronal markers or the feature indicating presence or absence of nervous systems in any evolutionary reconstruction because virtually the same genes are expressed in majority of non-neuronal tissues controlling cellular excitability in a broad spectrum of unicellular and multicellular eukaryotes. Not surprisingly, we found that ctenophores reveal examples both gene gain and loss compare, although ion channel complement is also reduced compared to eumetazoans.

If neurons evolved more than once, as our combined analysis and phylogeny imply, then what is a neuron? We define *neurons as polarized secretory cells specialized for directional active conducting – features that enable them to transmit signals, primarily chemical in nature, beyond their immediate neighbours without affecting all intervening cells en route.* As such multiple origins of neurons from different classes of secretory cells might occur more than once during ~ 600 million years of animal evolution. In the 1950-60s Grundfest and in 1970s Sakharov suggested that neurons arose from ancestral secretory cells, when the secretory activity became confined to the termination of elongated processes[152]. Early neurons/synapses evolved as the next step in development of compartmentalized transmitter secretion – the hallmark of neuronal organization - recruiting pre-existing molecular components for polarized transport and signalling from secretory and receptor machinery already well-developed in unicellular eukaryotes. This explains the recruitment in ctenophores of certain RNA binding proteins, acting as a cargo to transport selected localized RNA (e.g. secretory apparatus, receptors, ion channels, etc.) to distant neural processes. The first neural circuits evolved from undifferentiated secretory-like cells (perhaps without recognized *bona fide* neurons as in *Trichoplax*) to control cilia and coordinate primary (ciliated) locomotion[299] (e.g. using volume transmission vs localized synapses) whereas the first muscles controlled the hydroskeleton and feeding/defensive movements as in extant ctenophores.

What were the early transmitters? L-glutamate might be one of the first to be recruited as a neuromuscular transmitter in the ctenophores following profound diversification of iGluRs, components of Glu synthesis and uptake. A diversity of secretory signal peptides and their receptors (including expansion of ASICs) can also be recruited for this role in ctenophores independently from other metazoans, paralleled by the diversification of gap junction proteins most profoundly expressed in the apical organ of extant ctenophores. Polarized secretory (possible peptidergic) cells were likely involved in coordination of ciliated locomotion in many early animals and these types of cells can be considered as evolutionary precursors of different neuronal lineages with specific transmitter phenotypes. Our data also imply that classical low molecular weight transmitter systems such as cholinergic, GABAergic and three classes of monoaminergic systems (serotoninergic, histaminergic and catecholaminergic) were recruited for neuronal functions in cnidarian/bilaterians lineages. The presence of two existing neural nets in

ctenophores with ectodermal and mesogleal neurons that are similar to mesogleal muscle-like precursors raises the hypothesis that some neuronal lineages evolved from muscle cells that lost contractility and gained a polarized secretory/synaptic apparatus.

In conclusion, our genomic, expression and microchemical data suggest that the overall molecular makeup of the ctenophore nervous systems is remarkably different from all other nervous systems studied suggesting extensive parallel evolution of neural organization in this lineage. Regardless of evolutionary interpretations, the sequenced *Pleurobrachia* genome, combined genomic, metabolomic and physiological data on 10 different species revealed extraordinary and unique molecular diversity of developmental and neural signalling pathways. Together with numerous ctenophore "innovations", ctenophores may serve as a model to understand the origins/emergence of complex integrative functions and can be used in synthetic biology and regenerative medicine to design novel regulatory systems.

Summary

Ctenophores, or comb jellies, are enigmatic animals possibly tracing their ancestry to the Ediacaran biota. In contrast to other lineages, ctenophores have two distinct nervous systems; an "elementary brain", and "true" mesoderm-derived muscles supporting their predatory lifestyle. As such, they are imperative for understanding evolution of animal innovations, including origins of neural and muscular tissues. We present the first ctenophore genome, *Pleurobrachia bachei*, the most distinct from other animal genomes sequenced. Our data place Ctenophora as the earliest lineage within Metazoa, their unique organization and a possibility of independent development of both neuronal[152,229,295,296,300] and muscular[301] systems supporting their phenotypic complexity. Our conclusions are supported by combined metabolomics and functional data suggesting that L-glutamate[302] and numerous peptides might be the first transmitters in early nervous systems. These findings might have implementations for regenerative biomedicine and synthetic biology in designing novel signaling pathways and systems. In this case, ctenophores, and their genomes present matchless examples of "experiments" in nature (we might call them "*aliens of the sea*") and the possible preservation of ancient molecular toolkits lost in other animal lineages.

References for the Supplementary Data:

1 Kohn, A. B., Moroz, T. P., Barnes, J. P., Netherton, M. & Moroz, L. L. Single-cell semiconductor sequencing. *Methods Mol Biol* **1048**, 247-284, doi:10.1007/978-1-62703-556-9_18 (2013).

2 Moroz, L. L. & Kohn, A. B. Single-neuron transcriptome and methylome sequencing for epigenomic analysis of aging. *Methods Mol Biol* **1048**, 323-352, doi:10.1007/978-1-62703-556-9_21 (2013).

3 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829, doi:10.1101/gr.074492.107 gr.074492.107 [pii] (2008).

4 Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1:18**, 1-6 (2012).

5 Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**, 1117-1123, doi:10.1101/gr.089532.108 gr.089532.108 [pii] (2009).

6 Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-644, doi:10.1093/bioinformatics/btn013 btn013 [pii] (2008).

7 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 nmeth.1923 [pii] (2012).

8 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402, doi:gka562 [pii] (1997).

9 Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B. & Brom, T. H. in *arXiv:1203.4802 [q-bio.GN]*.

10 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664, doi:10.1101/gr.229202. Article published online before March 2002 (2002).

11 Srivastava, M. *et al.* The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* **466**, 720-726, doi:nature09201 [pii] 10.1038/nature09201 (2010).

12 Chapman, J. A. *et al.* The dynamic genome of *Hydra*. *Nature* **464**, 592-596, doi:nature08830 [pii] 10.1038/nature08830 (2010).

13 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 btp352 [pii] (2009).

14 Salamov, A. A. & Solovyev, V. V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res* **10**, 516-522 (2000).

15 Solovyev, V. in *Handbook of Statistical Genetics* (eds David J. Balding, Martin Bishop, & Chris Cannings) 97-159 (John Wiley & Sons, 2007).

16 Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **37**, D5-15, doi:gkn741 [pii] 10.1093/nar/gkn741 (2009).

17 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-467, doi:CGR20051101_4462 [pii] 10.1159/000084979 (2005).

18 Kapitonov, V. V., Tempel, S. & Jurka, J. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* **448**, 207-213, doi:S0378-1119(09)00416-8 [pii] 10.1016/j.gene.2009.07.019 (2009).

19 Girardo, D. O., Citarella, M., Kohn, A. B. & Moroz, L. L. Zero-click, automatic assembly, annotation and visualization workflow for comparative analysis of

transcriptomes: The quest for novel signaling pathways. *Integrative and Comparative Biology* 11.6 Abstract (2013).

20 Girardo, D. O., Citarella, M. R., Kohn, A. B. & Moroz, L. L. Automatic Transcriptome Analysis and Quest for Signaling Molecules In Basal Metazoans. *Integrative and Comparative Biology* Abstact meeting, P1.136 (2012).

21 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652, doi:10.1038/nbt.1883 nbt.1883 [pii] (2011).

22 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* **17**, 10-12 (2011).

23 Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res* **32**, D138-141, doi:10.1093/nar/gkh121 32/suppl_1/D138 [pii] (2004).

24 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

25 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30, doi:gkd027 [pii] (2000).

26 Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109-114, doi:gkr988 [pii] 10.1093/nar/gkr988 (2012).

27 Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-795, doi:10.1016/j.jmb.2004.05.028 S0022283604005972 [pii] (2004).

28 Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* **300**, 1005-1016, doi:10.1006/jmbi.2000.3903 S0022-2836(00)93903-2 [pii] (2000).

29 Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580, doi:10.1006/jmbi.2000.4315 S0022-2836(00)94315-8 [pii] (2001).

30 Kall, L., Krogh, A. & Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**, 1027-1036, doi:10.1016/j.jmb.2004.03.016 S0022283604002943 [pii] (2004).

31 Kocot, K. M. *et al.* Phylogenomics reveals deep molluscan relationships. *Nature* **477**, 452-456, doi:10.1038/nature10382 nature10382 [pii] (2011).

32 Lottaz, C., Iseli, C., Jongeneel, C. V. & Bucher, P. Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics* **19 Suppl 2**, ii103-112 (2003).

33 Ebersberger, I., Strauss, S. & Von Haeseler, A. HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology* **9**, 157 (2009).

34 Katoh, K., Kuma, K.-i., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research* **33**, 511 -518 %U http://nar.oxfordjournals.org/content/533/512/511.abstract (2005).

35 Roure, B., Rodriguez-Ezpeleta, N. & Philippe, H. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol Biol* **7 Suppl 1**, S2, doi:1471-2148-7-S1-S2 [pii] 10.1186/1471-2148-7-S1-S2 (2007).

36 Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564-577, doi:780704285 [pii] 10.1080/10635150701472164 (2007).

37 Misof, B. & Misof, K. A Monte Carlo Approach Successfully Identifies Randomness in Multiple Sequence Alignments : A More Objective Means of Data Exclusion. *Systematic Biology* **58**, 21 -34 %U http://sysbio.oxfordjournals.org/content/58/21/21.abstract (2009).
38 Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745-749, doi:nature06614 [pii] 10.1038/nature06614 (2008).
39 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690, doi:btl446 [pii] 10.1093/bioinformatics/btl446 (2006).
40 Smith, S. A. & Dunn, C. W. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24**, 715 -716 %U http://bioinformatics.oxfordjournals.org/content/724/715/715.abstract (2008).
41 Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B. & Brom, T. H. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data.:.. . *arXiv* **1203**, 4802 (2012).
42 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:btm404 [pii] 10.1093/bioinformatics/btm404 (2007).
43 Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31**, 3497-3500 (2003).
44 Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. Multiple sequence alignment with Clustal X. *Trends Biochem Sci* **23**, 403-405 (1998).
45 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113, doi:10.1186/1471-2105-5-113 1471-2105-5-113 [pii] (2004).
46 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-552 (2000).
47 Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731-2739, doi:msr121 [pii] 10.1093/molbev/msr121 (2011).
48 Ptitsyn, A. & Moroz, L. L. Computational workflow for analysis of gain and loss of genes in distantly related genomes. *BMC Bioinformatics* **13 Suppl 15**, S5, doi:10.1186/1471-2105-13-S15-S5 1471-2105-13-S15-S5 [pii] (2012).
49 Derelle, R. & Lang, B. F. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol Biol Evol* **29**, 1277-1289, doi:10.1093/molbev/msr295 msr295 [pii] (2012).
50 Hampl, V. *et al.* Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proc Natl Acad Sci U S A* **106**, 3859-3864, doi:10.1073/pnas.0807880106 0807880106 [pii] (2009).
51 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797, doi:10.1093/nar/gkh340 32/5/1792 [pii] (2004).
52 de Vienne, D. M., Ollier, S. & Aguileta, G. Phylo-MCOA: a fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. *Mol Biol Evol* **29**, 1587-1598, doi:10.1093/molbev/msr317 msr317 [pii] (2012).
53 Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286-2288, doi:10.1093/bioinformatics/btp368 btp368 [pii] (2009).

54 Huerta-Cepas, J., Dopazo, J. & Gabaldon, T. ETE: a python Environment for Tree Exploration. *BMC Bioinformatics* **11**, 24, doi:10.1186/1471-2105-11-24 1471-2105-11-24 [pii] (2010).
55 Ryan, J. F. *et al.* The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. *Science* **342**, 1242592, doi:10.1126/science.1242592 342/6164/1242592 [pii] (2013).
56 Osigus, H. J., Eitel, M., Bernt, M., Donath, A. & Schierwater, B. Mitogenomics at the base of Metazoa. *Mol Phylogenet Evol* **69**, 339-351, doi:10.1016/j.ympev.2013.07.016 S1055-7903(13)00293-5 [pii] (2013).
57 Dabe, E. C. *et al.* Epigenomic Signatures in Basal Metazoans: DNA Methyltransferase in Pleurobrachia bachei. *Integrative and Comparative Biology* **P1.151 Abstract** (2013).
58 Derelle, R., Lopez, P., Le Guyader, H. & Manuel, M. Homeodomain proteins belong to the ancestral molecular toolkit of eukaryotes. *Evol Dev* **9**, 212-219, doi:EDE153 [pii] 10.1111/j.1525-142X.2007.00153.x (2007).
59 Jezzini, S. H., Bodnarova, M. & Moroz, L. L. Two-color in situ hybridization in the CNS of *Aplysia californica*. *J Neurosci Methods* **149**, 15-25 (2005).
60 Moroz, L. L. *et al.* Neuronal transcriptome of *Aplysia*: neuronal compartments and circuitry. *Cell* **127**, 1453-1467 (2006).
61 Moroz, L. L. & Kohn, A. B. Analysis of gene expression in neurons and synapses by multi-color *in situ* hybridization *Methods in Molecular Biology*, in press (2013).
62 Fuller, R. R., Moroz, L. L., Gillette, R. & Sweedler, J. V. Single neuron analysis by capillary electrophoresis with fluorescence spectroscopy. *Neuron* **20**, 173-181, doi:S0896-6273(00)80446-8 [pii] (1998).
63 Lapainis, T., Rubakhin, S. S. & Sweedler, J. V. Capillary electrophoresis with electrospray ionization mass spectrometric detection for single-cell metabolomics. *Anal Chem* **81**, 5858-5864, doi:10.1021/ac900936g (2009).
64 Nemes, P., Knolhoff, A. M., Rubakhin, S. S. & Sweedler, J. V. Metabolic differentiation of neuronal phenotypes by single-cell capillary electrophoresis-electrospray ionization-mass spectrometry. *Anal Chem* **83**, 6810-6817, doi:10.1021/ac2015855 (2011).
65 Moroz, L. L., Dahlgren, R. L., Boudko, D., Sweedler, J. V. & Lovell, P. Direct single cell determination of nitric oxide synthase related metabolites in identified nitrergic neurons. *J Inorg Biochem* **99**, 929-939, doi:S0162-0134(05)00015-2 [pii] 10.1016/j.jinorgbio.2005.01.013 (2005).
66 Cohen, N. M., Kenigsberg, E. & Tanay, A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* **145**, 773-786, doi:10.1016/j.cell.2011.04.024 S0092-8674(11)00482-X [pii] (2011).
67 Shinzato, C. *et al.* Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature* **476**, 320-323, doi:10.1038/nature10249 nature10249 [pii] (2011).
68 Putnam, N. H. *et al.* Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86-94, doi:317/5834/86 [pii] 10.1126/science.1139158 (2007).
69 Srivastava, M. *et al.* The *Trichoplax* genome and the nature of placozoans. *Nature* **454**, 955-960, doi:nature07191 [pii] 10.1038/nature07191 (2008).
70 Jex, A. R. *et al. Ascaris suum* draft genome. *Nature* **479**, 529-533, doi:10.1038/nature10553 nature10553 [pii] (2011).

71 Zhang, G. *et al.* The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**, 49-54, doi:10.1038/nature11413 nature11413 [pii] (2012).
72 Sodergren, E. *et al.* The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**, 941-952, doi:314/5801/941 [pii] 10.1126/science.1133609 (2006).
73 Davidson, C. R., Best, N. M., Francis, J. W., Cooper, E. L. & Wood, T. C. Toll-like receptor genes (TLRs) from *Capitella capitata* and *Helobdella robusta* (Annelida). *Dev Comp Immunol* **32**, 608-612, doi:10.1016/j.dci.2007.11.004 S0145-305X(07)00205-4 [pii] (2008).
74 Augustin, R., Fraune, S. & Bosch, T. C. How *Hydra* senses and destroys microbes. *Semin Immunol* **22**, 54-58, doi:10.1016/j.smim.2009.11.002 S1044-5323(09)00116-X [pii] (2010).
75 Gauthier, M. E., Du Pasquier, L. & Degnan, B. M. The genome of the sponge *Amphimedon queenslandica* provides new perspectives into the origin of Toll-like and interleukin 1 receptor pathways. *Evol Dev* **12**, 519-533, doi:10.1111/j.1525-142X.2010.00436.x (2010).
76 Kerner, P., Degnan, S. M., Marchand, L., Degnan, B. M. & Vervoort, M. Evolution of RNA-binding proteins in animals: insights from genome-wide analysis in the sponge *Amphimedon queenslandica*. *Mol Biol Evol* **28**, 2289-2303, doi:msr046 [pii] 10.1093/molbev/msr046 (2011).
77 Nielsen, C. *Animal Evolution: Interrelationships of the living phyla*. 402 (Oxford University Press, 2012).
78 Hernandez-Nicaise, M.-L. in *Microscopic Anatomy of Invertebrates: Placozoa, Porifera, Cnidaria, and Ctenophora* Vol. 2 (eds F. W. F.W. Harrison & J. A. Westfall) 359-418 (Wiley, 1991).
79 Tamm, S. L. in *Electrical conduction and behavior in "simple" invertebrates* 266-358 (Clarendon Press, 1982).
80 Ryan, J. F., Pang, K., Mullikin, J. C., Martindale, M. Q. & Baxevanis, A. D. The homeodomain complement of the ctenophore *Mnemiopsis leidyi* suggests that Ctenophora and Porifera diverged prior to the ParaHoxozoa. *Evodevo* **1**, 9, doi:2041-9139-1-9 [pii] 10.1186/2041-9139-1-9 (2010).
81 Jager, M. *et al.* New insights on ctenophore neural anatomy: immunofluorescence study in *Pleurobrachia pileus* (Muller, 1776). *J Exp Zool B Mol Dev Evol* **316B**, 171-187, doi:10.1002/jez.b.21386 (2011).
82 Horridge, G. A. Recent studeis on the Ctenophora. *Coelenterate Biology: Reviews and New Perspectives*, 439-468 (1974).
83 Tamm, S. L. Mechanisms of ciliary co-ordination in ctenophores. *Journal Experimental Biology* **59**, 231-245 (1973).
84 Tamm, S. L. & Moss, A. G. Unilateral ciliary reversal and motor responses during prey capture by the ctenophore *Pleurobrachia*. *Journal Experimental Biology* **144**, 443-461 (1985).
85 Jager, M. R., Chiori, A., Dayraud, A. C., Queinnec, E. & Manuel, M. New insights on ctenophore neural anatomy: immunofluorescence study in *Pleurobrachia pileus* (Mueller, 1776). *Journal Experimental Zoology (Molecular Developmental Evolution)* **314B**, 1-17 (2010).
86 Hyman, L. H. *The Invertebrates*. Vol. Volume I: Protozoas through Ctenophora 726 (McGraw-Hill Book Company).

87 Greene, C. H. Patterns of prey selection: implications of predator foraging tactics. *American Naturalist* **128**, 824-829 (1986).
88 von Byern, J., Mills, C. E. & Flammang, P. in *Biological Adhesive Systems: From Nature to Technical and Medical Application* 22-29 (Springer Publishing Company, 2010).
89 Hirota, J. Quantitative natural history of *Pleurobrachia bachei*. *La Jolla Bight. Fishery Bulletin* **72**, 295-335 (1974).
90 Haddock, S. H. D. Not all ctenophores are bioluminescent: *Pleurobrachia*. *Biological Bulletin* **189**, 356-362 (1995).
91 H., D. P. A.C. Geiese and J.S. Pearse, eds. edn, Vol. 1 Acoelomate and Pseudocoelomate Metazoans 201-265 (1974).
92 Dunlap, H. L. Oogenesis in the Ctenophora. *Dissertation, University of Washington Seattle* (1966).
93 J., H. Laboratory culture and metabolism of the planktonic ctenophore *Pleurobrachia bachei*. *Biological Oceanography of the Northern North Pacific Ocean*, 465-484 (1972).
94 Wrobel, D. & Mills, C. *Pacific Coast Pelagic Invertebrates: a Guide to the Common Gelatinous Animals*. Vol. iv (Sea Challengers and the Monterey Bay Aquarium, 1998).
95 Mills, C. E. Patterns and Mechanisms of Vertical Distribution of Medusae and Ctenophores. *PhD Dissertation University of Victoria British Columbia, CA*, 384 pgs (1982).
96 Mackie, G. O. & Mills, C. E. Use of the *Pisces IV* submersible for zooplankton studies in coastal waters of British Columbia. *Canadian Journal of Fish and Aquatic Science* **40**, 763-776 (1983).
97 Agassiz, A. & Agassiz, G. R. *Letters and recollections of Alexander Agassiz, with a sketch of his life and work*. (Houghton Mifflin Company, 1913).
98 Agassiz, L. Vol. 3 301 and 303 plates (Little Brown and Co., Boston, MA, 1860).
99 Agassiz, A. in *Order Ctenophorae Esch. Mem. Mus.* Vol. 1 7-40 (Comparative Zoology Harvard College, 1865).
100 Chun, C. in *Fauna und Flora des Golfes von Neapel, herausgegeben von der zoologischen Station in Neapel.* Vol. I. Monographie XVIII. 313 pages with 318 pages, 322 text-figures (1880).
101 Moser, F. Die Ctenophoren der Deutschen Südpolar-Expedition. *Zoologie* **11**, 117-192 plates 120-122 (1909).
102 Bigelow, H. B. Report XXVI: The ctenophores. Reports on the scientific results of the Expedition to the Eastern Tropical Pacific, in charge of Alexander Agassiz, by the U.S. Fish Commission Steamer "Albatross," from October, 1904, to March, 1905, Lieut. Commander L. M. Garrett, U. S. N., Commanding. *Comanding Bull. Mus Comp Zoology* **54**, 369-404 plates 361-362 (1912).
103 Mayer, A. G. Ctenophores of the Atlantic Coast of North America. *Carnegie Institution of Washington, Publication No.* **162**, 1-58 (1912).
104 Ralph, P. M. & Kaberry, C. Ctenophores from the waters of Cook Strait and Wellington Harbour. *Zoology Publication Victoria University College* **No 3**, 11 pp (1950).
105 Tregouboff, G. & Rose, M. Vol. I & II 589 pg, 207 plates (Centre national de la recherche scientifique, Paris, 1957).
106 Palma, S. G. & Meruane, J. Z. Aspectos ecologicos y crecimiento de *Pleurobrachia pileus* (Ctenophora) en la region de Valparaiso. *Investigative Marine Valparaiso* **6**, 25-40 (1975).

107 Mills, C. E. in *Marine Invertebrates of he Pacific Northwest* Vol. Seattle (ed E.N. Kolzloff & L.H. rice) 79-81 (University of Washington Press, 1987a).
108 Mills, C. E. Revised classification of the genus *Euplokamis* Chun, 1880 (Ctenophora: Cydippida: Euplokamidae n. fam.) with a description of the new species *Euplokamis dunlapae*. *Canadian Journal Zoology* **65**, 2661-2668 (1987b).
109 Vinogradov, M. E., Arashkevich, E. G. & Ilchenko, S. V. The ecology of the *Calanus ponticus* population in the deeper layer of its concentration in the Black Sea. *Journal Planton Research* **14**, 443-461 (1992).
110 Mills, C. E., Pugh, P. R., Harbison, G. R. & Haddock, S. H. D. Medusae, siphonophores and ctenophores of the Alboran Sea, south west Mediterranean. *Scientia Marina* **60**, 145-163 (1996).
111 Buecher, E. & Gasser, B. Estimation of predatory impact of *Pleurobrachia rhodopsis* (cydippid ctenophore) in the northwestern Mediterranean Sea: *in situ* observations and laboratory experiments. *Journal of Plankton Research* **20**, 631-651 (1998).
112 Gibbons, M. J., Buecher, E. & Thibault-Botha, D. Observations on the ecology of *Pleurobrachia pileus* (Ctenophora) in the Southern Benguela ecosystem *African Journal Marine Science* **25**, 253-261 (2003).
113 Freeman, G. The establishment of the oral-aboral axis in the ctenophore embryo. *Journal Embryology Experimental Morphology* **42**, 237-260 (1977).
114 Mills, C. E. & Haddock, S. H. D. in *Light and Smith's Manual: Intertidal Invertebrates of the Central California Coast* (ed J.T. Carlton) 189-199 (University of California Press, 2007).
115 Mills, C. E. in *Biology Marine Mediterranian* (ed G. Relini) Ch. Ctenophora, 102-104 (2008).
116 Molinero, J. C. *et al.* Climate control on the long-term anomalous changes of zooplankton communities in the Northwestern Mediterranean. *Global Change Biology* **14**, 11-26 (2008).
117 Yanez, E. *et al.* Seamounts in the southeastern Pacific Ocean and biodiversity on Juan Fernandez seamounts, Chile. *Latin American Journal of Aquatic Research* **37**, 555-570 (2009).
118 Mianzan, H., Dawson, E. W. & Mills, C. E. in *New Zealand Inventory of Biodiversity* Vol. Volume One: Kingdom Animalia: Radiata, Lophotrochozoa, and Deuterostomia (ed D.P. Gordon) 49-58 (Canterbury University Press, Christchurch, 2009).
119 Mianzan, H. in *South Atlantic Zooplankton* 561-573 (Backhuys Publishers, 1999).
120 Hernandez-Nicaise, M. L. The nervous system of ctenophores. III Ultrastructure of synapses. *Journal Neurocytology* **2**, 249-263 (1973).
121 Franc, J. M. Organization and function of ctenophore colloblasts: an ultrastructural study. *Biological Bulletin* **155**, 527-541 (1978).
122 Laval, P. *et al.* Small-scale distribution of macroplankton and micronekton in the Ligurian Sea (Mediterranean Sea) as observed from the manned submersible *Cyana*. *Journal Plankton Research* **11**, 665-685 (1989).
123 Jager, M., Queinnec, E., Houliston, E. & Manuel, M. Expansion of the *SOX* gene family predated the emergence of the Bilateria. *Molecular Phylogenetics Evolution* **39**, 468-477 (2006).

124 Jager, M., Queinnec, E., Chiori, R., Le Guyader, H. & Manuel, M. Insights into the early evolution of the SOX genes from expression analyses in a ctenophore. *Journal of Experimental Zoology (Molecular Developmental Evolution)* **310B**, 650-667 (2008).
125 Derelle, R. T., Momose, M., Manuel, C., Da Silva, P. & Houliston, W. E. Convergent origins and rapid evolution of spliced leader trans-splicing in Metazoa: insights from the Cteenophora and Hydrozoa. *RNA* **16**, 696-707 (2010).
126 Alie, A. *et al.* Somatic stem cells express Piwi and Vasa genes in an adult ctenophore: Ancient association of "germline genes" with stemness. *Developmental Biology* **350**, 183-197, doi:10.1016/j.ydbio.2010.10.019 (2011).
127 Kovalev, A. V. & Piontkovski, S. A. Interannual changes in the biomass of the Black Sea gelatinous zooplankton. *Journal Plankton Research* **20**, 1377-1385 (1998).
128 Petran, A. & Moldoveanu, M. Post-invasion ecological impact of the Atlantic ctenophore *Mnemiopsis leidyi* Agassiz, 1865 on the zooplankton from the Romanian Black Sea waters. *Cercetari Marine* **135-137**, 135-157 (1994-1995).
129 Bayha, K., Harbison, G. R., McDonald, J. & Gaffney, P. K. in *Aquatic Invasions in the Black, Caspian, and Mediterranean Sea* Vol. part 1 *NATO Science Series: IV: Earth and Environmental Sciences* 167-175 (2004).
130 Jurka, J., Kapitonov, V. V., Kohany, O. & Jurka, M. V. Repetitive sequences in complex genomes: structure and evolution. *Annu Rev Genomics Hum Genet* **8**, 241-259, doi:10.1146/annurev.genom.8.080706.092416 (2007).
131 Putnam, N. H. *et al.* The Amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064-1071, doi:nature06967 [pii] 10.1038/nature06967 (2008).
132 Kapitonov, V. V. & Jurka, J. Molecular paleontology of transposable elements from *Arabidopsis thaliana*. *Genetica* **107**, 27-37 (1999).
133 Kapitonov, V. V. & Jurka, J. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *Proc Natl Acad Sci U S A* **100**, 6569-6574, doi:10.1073/pnas.0732024100 0732024100 [pii] (2003).
134 Bird, A. *et al.* Studies of DNA methylation in animals. *J Cell Sci Suppl* **19**, 37-39 (1995).
135 Henderson, I. R. & Jacobsen, S. E. Epigenetic inheritance in plants. *Nature* **447**, 418-424 (2007).
136 Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet* **14**, 9-25 (1975).
137 Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187**, 226-232 (1975).
138 Jahner, D. *et al.* De novo methylation and expression of retroviral genomes during mouse embryogenesis. *Nature* **298**, 623-628 (1982).
139 Stewart, C. L., Stuhlmann, H., Jahner, D. & Jaenisch, R. De novo methylation, expression, and infectivity of retroviral genomes introduced into embryonal carcinoma cells. *Proc Natl Acad Sci U S A* **79**, 4098-4102 (1982).
140 Jaenisch, R., Harbers, K., Jahner, D., Stewart, C. & Stuhlmann, H. DNA methylation, retroviruses, and embryogenesis. *J Cell Biochem* **20**, 331-336 (1982).
141 Tweedie, S., Charlton, J., Clark, V. & Bird, A. Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Mol Cell Biol* **17**, 1469-1475 (1997).
142 Allis, C. D., Jenuwein, T.,Reinberg, D., and Caparros, M.L., eds. *Epigenetics*. ((Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press), 2007).

143 Goll, M. G. & Bestor, T. H. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* **74**, 481-514, doi:10.1146/annurev.biochem.74.010904.153721 (2005).
144 Goll, M. G. *et al.* Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science* **311**, 395-398 (2006).
145 Albalat, R. Evolution of DNA-methylation machinery: DNA methyltransferases and methyl-DNA binding proteins in the amphioxus *Branchiostoma floridae*. *Dev Genes Evol* **218**, 691-701, doi:10.1007/s00427-008-0247-7 (2008).
146 Martienssen, R. A. & Colot, V. DNA methylation and epigenetic inheritance in plants and filamentous fungi. *Science* **293**, 1070-1074, doi:293/5532/1070 [pii] 10.1126/science.293.5532.1070 (2001).
147 Denis, H., Ndlovu, M. N. & Fuks, F. Regulation of mammalian DNA methyltransferases: a route to new mechanisms. *EMBO Rep* **12**, 647-656, doi:embor2011110 [pii] 10.1038/embor.2011.110 (2011).
148 Song, C. X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* **29**, 68-72, doi:nbt.1732 [pii] 10.1038/nbt.1732 (2011).
149 Rein, T., DePamphilis, M. L. & Zorbas, H. Identifying 5-methylcytosine and related modifications in DNA genomes. *Nucleic Acids Res* **26**, 2255-2264 (1998).
150 Zilberman, D. & Henikoff, S. Genome-wide analysis of DNA methylation patterns. *Development* **134**, 3959-3965 (2007).
151 Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-935, doi:1170116 [pii] 10.1126/science.1170116 (2009).
152 Moroz, L. L. On the independent origins of complex brains and neurons. *Brain, behavior and evolution* **74**, 177-190 (2009).
153 Nosenko, T. *et al.* Deep metazoan phylogeny: When different genes tell different stories. *Mol Phylogenet Evol* **67**, 223-233, doi:10.1016/j.ympev.2013.01.010 S1055-7903(13)00029-8 [pii] (2013).
154 Pick, K. S. *et al.* Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol Biol Evol* **27**, 1983-1987, doi:msq089 [pii] 10.1093/molbev/msq089 (2010).
155 Bourlat, S. J. *et al.* Deuterostome phylogeny reveals monophyletic chordates and the new phylum *Xenoturbellida*. *Nature* **444**, 85-88 (2006).
156 Hejnol, A. *et al.* Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B: Biological Sciences* **276**, 4261 (2009).
157 Philippe, H. *et al.* Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* **19**, 706-712, doi:S0960-9822(09)00805-7 [pii] 10.1016/j.cub.2009.02.052 (2009).
158 Berezikov, E. Evolution of microRNA diversity and regulation in animals. *Nat Rev Genet* **12**, 846-860, doi:10.1038/nrg3079 nrg3079 [pii] (2011).
159 Mikhailov, K. V. *et al.* The origin of Metazoa: a transition from temporal to spatial cell differentiation. *Bioessays* **31**, 758-768 (2009).
160 Erwin, D. H. *et al.* The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science* **334**, 1091-1097, doi:10.1126/science.1206375 334/6059/1091 [pii] (2011).

161 Lange, C. *et al.* Defining the origins of the NOD-like receptor system at the base of animal evolution. *Mol Biol Evol* **28**, 1687-1702 (2011).
162 Leulier, F. & Lemaitre, B. Toll-like receptors--taking an evolutionary approach. *Nat Rev Genet* **9**, 165-178, doi:nrg2303 [pii] 10.1038/nrg2303 (2008).
163 Steinmetz, P. R. *et al.* Independent evolution of striated muscles in cnidarians and bilaterians. *Nature* **487**, 231-234, doi:nature11180 [pii] 10.1038/nature11180 (2012).
164 Kusserow, A. *et al.* Unexpected complexity of the Wnt gene family in a sea anemone. *Nature* **433**, 156-160 (2005).
165 Lengfeld, T. *et al.* Multiple Wnts are involved in *Hydra* organizer formation and regeneration. *Dev Biol* **330**, 186-199, doi:10.1016/j.ydbio.2009.02.004 (2009).
166 Pang, K., Ryan, J. F., Baxevanis, A. D. & Martindale, M. Q. Evolution of the TGF-beta signaling pathway and its potential role in the ctenophore, *Mnemiopsis leidyi*. *PLoS One* **6**, e24152, doi:10.1371/journal.pone.0024152 PONE-D-11-09363 [pii] (2011).
167 Sebe-Pedros, A., de Mendoza, A., Lang, B. F., Degnan, B. M. & Ruiz-Trillo, I. Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsaspora owczarzaki*. *Mol Biol Evol* **28**, 1241-1254, doi:10.1093/molbev/msq309 (2011).
168 Suga, H. *et al.* The *Capsaspora* genome reveals a complex unicellular prehistory of animals. *Nat Commun* **4**, 2325, doi:10.1038/ncomms3325 ncomms3325 [pii] (2013).
169 Bürglin, T. R. *Pp. 25-72 in Double D, ed. A comprehensive classification of homeobox genes. Guidebook to the Homeobox genes. Oxford University Press., New York.*, (1994).
170 Holland, P. W., Booth, H. A. & Bruford, E. A. Classification and nomenclature of all human homeobox genes. *BMC Biol* **5**, 47, doi:1741-7007-5-47 [pii] 10.1186/1741-7007-5-47 (2007).
171 Larroux, C. *et al.* The NK homeobox gene cluster predates the origin of Hox genes. *Curr Biol* **17**, 706-710, doi:S0960-9822(07)01075-5 [pii] 10.1016/j.cub.2007.03.008 (2007).
172 Massari, M. E. & Murre, C. Helix-loop-helix proteins: regulators of transcription in eucaryotic organisms. *Mol Cell Biol* **20**, 429-440 (2000).
173 Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-6067, doi:10.1073/pnas.0400782101 (2004).
174 Mangelsdorf, D. J. *et al.* The nuclear receptor superfamily: the second decade. *Cell* **83**, 835-839 (1995).
175 Reitzel, A. M. & Tarrant, A. M. Nuclear receptor complement of the cnidarian *Nematostella vectensis*: phylogenetic relationships and developmental expression patterns. *BMC Evol Biol* **9**, 230, doi:10.1186/1471-2148-9-230 (2009).
176 Bridgham, J. T. *et al.* Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS Biol* **8**, doi:10.1371/journal.pbio.1000497 (2010).
177 Reitzel, A. M. *et al.* Nuclear receptors from the ctenophore *Mnemiopsis leidyi* lack a zinc-finger DNA-binding domain: lineage-specific loss or ancestral condition in the emergence of the nuclear receptor superfamily? *Evodevo* **2**, 3, doi:2041-9139-2-3 [pii] 10.1186/2041-9139-2-3 (2011).
178 Sackton, T. B. *et al.* Dynamic evolution of the innate immune system in Drosophila. *Nat Genet* **39**, 1461-1468 (2007).
179 Buckley, K. M. & Rast, J. P. Characterizing immune receptors from new genome sequences. *Methods Mol Biol* **748**, 273-298 (2011).

180 Seth, R. B., Sun, L., Ea, C. K. & Chen, Z. J. Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-kappaB and IRF 3. *Cell* **122**, 669-682, doi:S0092-8674(05)00816-0 [pii] 10.1016/j.cell.2005.08.012 (2005).
181 Meylan, E. *et al.* Cardif is an adaptor protein in the RIG-I antiviral pathway and is targeted by hepatitis C virus. *Nature* **437**, 1167-1172, doi:nature04193 [pii] 10.1038/nature04193 (2005).
182 Kato, H. *et al.* Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature* **441**, 101-105, doi:nature04734 [pii] 10.1038/nature04734 (2006).
183 Yoneyama, M. *et al.* Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity. *J Immunol* **175**, 2851-2858, doi:175/5/2851 [pii] (2005).
184 Jiang, F. *et al.* Structural basis of RNA recognition and activation by innate immune receptor RIG-I. *Nature* **479**, 423-427, doi:10.1038/nature10537 nature10537 [pii] (2011).
185 Kawai, T. *et al.* IPS-1, an adaptor triggering RIG-I- and Mda5-mediated type I interferon induction. *Nat Immunol* **6**, 981-988, doi:ni1243 [pii] 10.1038/ni1243 (2005).
186 Mukherjee, K., Korithoski, B. & Kolaczkowski, B. Ancient origins of vertebrate-specific innate antiviral immunity. *Mol Biol Evol*, in Press (2013).
187 Bertin, J. *et al.* Human CARD4 protein is a novel CED-4/Apaf-1 cell death family member that activates NF-kappaB. *J Biol Chem* **274**, 12955-12958 (1999).
188 Ghildiyal, M. & Zamore, P. D. Small silencing RNAs: an expanding universe. *Nat Rev Genet* **10**, 94-108, doi:nrg2504 [pii] 10.1038/nrg2504 (2009).
189 Jost, D., Nowojewski, A. & Levine, E. Small RNA biology is systems biology. *BMB Rep* **44**, 11-21, doi:10.5483/BMBRep.2011.44.1.11 (2011).
190 Moroz, L. L. *et al.* The Genome of the Ctenophore *Pleurobrachia bachei*:Molecular Insights into Independent Origins of Nervous Systems. *Integrative and Comparative Biology* **66.7** Abstract (2012).
191 Moroz, L. L. Genomic Bases for Independent Origins of Neurons and Complex Brains: New Insights from RNA-seq and genomic sequencing of basal metazoans, basal deuterostomes and molluscs. *Integrative and Comparative Biology* **10.5 Abstract** (2013).
192 Bostwick, C. J. *et al.* Piwi genes and their expression in the ctenophore *Pleurobrachia bachei*: Quest for ancestral master regulators of non-coding RNAs in animals. *Integrative and Comparative Biology* **P3.75** Abstract (2013).
193 Filipowicz, W. RNAi: the nuts and bolts of the RISC machine. *Cell* **122**, 17-20, doi:10.1016/j.cell.2005.06.023 (2005).
194 Kawamata, T. & Tomari, Y. Making RISC. *Trends Biochem Sci* **35**, 368-376, doi:S0968-0004(10)00063-0 [pii] 10.1016/j.tibs.2010.03.009 (2010).
195 Grimson, A. *et al.* Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455**, 1193-1197, doi:nature07415 [pii] 10.1038/nature07415 (2008).
196 Mukherjee, K., Campos, H. & Kolaczkowski, B. Evolution of animal and plant dicers: early parallel duplications and recurrent adaptation of antiviral RNA binding in plants. *Mol Biol Evol* **30**, 627-641, doi:10.1093/molbev/mss263 mss263 [pii] (2013).
197 Christodoulou, F. *et al.* Ancient animal microRNAs and the evolution of tissue identity. *Nature* **463**, 1084-1088, doi:nature08744 [pii] 10.1038/nature08744 (2010).
198 Berezikov, E. *et al.* Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* **38**, 1375-1377 (2006).

199 Berezikov, E. *et al.* Deep annotation of *Drosophila melanogaster* microRNAs yields insights into their processing, modification, and emergence. *Genome Res* **21**, 203-215, doi:gr.116657.110 [pii] 10.1101/gr.116657.110 (2011).
200 Berger, W., Steiner, E., Grusch, M., Elbling, L. & Micksche, M. Vaults and the major vault protein: novel roles in signal pathway regulation and immunity. *Cell Mol Life Sci* **66**, 43-61, doi:10.1007/s00018-008-8364-z (2009).
201 Correa, R. L., Steiner, F. A., Berezikov, E. & Ketting, R. F. MicroRNA-directed siRNA biogenesis in *Caenorhabditis elegans*. *PLoS Genet* **6**, e1000903, doi:10.1371/journal.pgen.1000903 (2010).
202 Maxwell, E. K., Ryan, J. F., Schnitzler, C. E., Browne, W. E. & Baxevanis, A. D. MicroRNAs and essential components of the microRNA processing machinery are not encoded in the genome of the ctenophore Mnemiopsis leidyi. *BMC Genomics* **13**, 714, doi:10.1186/1471-2164-13-714 (2012).
203 de Jong, D. *et al.* Multiple dicer genes in the early-diverging metazoa. *Mol Biol Evol* **26**, 1333-1340, doi:msp042 [pii] 10.1093/molbev/msp042 (2009).
204 Garrett, S. & Rosenthal, J. J. RNA editing underlies temperature adaptation in K+ channels from polar octopuses. *Science* **335**, 848-851, doi:10.1126/science.1212795 science.1212795 [pii] (2012).
205 Knoop, V. When you can't trust the DNA: RNA editing changes transcript sequences. *Cell Mol Life Sci* **68**, 567-586, doi:10.1007/s00018-010-0538-9 (2011).
206 Gray, M. W. Evolutionary origin of RNA editing. *Biochemistry* **51**, 5235-5242, doi:10.1021/bi300419r (2012).
207 Benne, R. *et al.* Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46**, 819-826, doi:0092-8674(86)90063-2 [pii] (1986).
208 Chester, A. *et al.* The apolipoprotein B mRNA editing complex performs a multifunctional cycle and suppresses nonsense-mediated decay. *EMBO J* **22**, 3971-3982, doi:10.1093/emboj/cdg369 (2003).
209 Conticello, S. G. The AID/APOBEC family of nucleic acid mutators. *Genome Biol* **9**, 229, doi:gb-2008-9-6-229 [pii] 10.1186/gb-2008-9-6-229 (2008).
210 Nishikura, K. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* **79**, 321-349, doi:10.1146/annurev-biochem-060208-105251 (2010).
211 Wulff, B. E. & Nishikura, K. Substitutional A-to-I RNA editing. *Wiley Interdiscip Rev RNA* **1**, 90-101, doi:10.1002/wrna.10 (2010).
212 Sommer, B., Kohler, M., Sprengel, R. & Seeburg, P. H. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* **67**, 11-19, doi:0092-8674(91)90568-J [pii] (1991).
213 Burns, C. M. *et al.* Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature* **387**, 303-308, doi:10.1038/387303a0 (1997).
214 Gerber, A. P. & Keller, W. RNA editing by base deamination: more enzymes, more targets, new mysteries. *Trends Biochem Sci* **26**, 376-384, doi:S0968-0004(01)01827-8 [pii] (2001).
215 Tonkin, L. A. *et al.* RNA editing by ADARs is important for normal behavior in *Caenorhabditis elegans*. *EMBO J* **21**, 6025-6035 (2002).

216 Keegan, L. P., Leroy, A., Sproul, D. & O'Connell, M. A. Adenosine deaminases acting on RNA (ADARs): RNA-editing enzymes. *Genome Biol* **5**, 209, doi:10.1186/gb-2004-5-2-209 gb-2004-5-2-209 [pii] (2004).
217 Kawahara, Y. & Nishikura, K. Extensive adenosine-to-inosine editing detected in Alu repeats of antisense RNAs reveals scarcity of sense-antisense duplex formation. *FEBS Lett* **580**, 2301-2305, doi:S0014-5793(06)00357-7 [pii] 10.1016/j.febslet.2006.03.042 (2006).
218 Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol* **8**, 479-490, doi:nrm2178 [pii] 10.1038/nrm2178 (2007).
219 Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* **582**, 1977-1986, doi:S0014-5793(08)00207-X [pii] 10.1016/j.febslet.2008.03.004 (2008).
220 Musunuru, K. & Darnell, R. B. Determination and augmentation of RNA sequence specificity of the Nova K-homology domains. *Nucleic Acids Res* **32**, 4852-4861, doi:10.1093/nar/gkh799 32/16/4852 [pii] (2004).
221 Darnell, R. B. & Posner, J. B. Paraneoplastic syndromes involving the nervous system. *N Engl J Med* **349**, 1543-1554, doi:10.1056/NEJMra023009 349/16/1543 [pii] (2003).
222 Buckanovich, R. J. & Darnell, R. B. The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. *Mol Cell Biol* **17**, 3194-3201 (1997).
223 Darnell, R. B. Developing global insight into RNA regulation. *Cold Spring Harb Symp Quant Biol* **71**, 321-327, doi:10.1101/sqb.2006.71.002 (2006).
224 Lubin, F. D., Roth, T. L. & Sweatt, J. D. Epigenetic regulation of BDNF gene transcription in the consolidation of fear memory. *J Neurosci* **28**, 10576-10586, doi:28/42/10576 [pii] 10.1523/JNEUROSCI.1786-08.2008 (2008).
225 Colonques, J., Ceron, J., Reichert, H. & Tejedor, F. J. A transient expression of Prospero promotes cell cycle exit of *Drosophila* postembryonic neurons through the regulation of Dacapo. *PLoS One* **6**, e19342, doi:10.1371/journal.pone.0019342 (2011).
226 Satoh, G., Wang, Y., Zhang, P. & Satoh, N. Early development of Amphioxus nervous system with special reference to segmental cell organization and putative sensory cell precursors: a study based on the expression of pan-neuronal marker gene Hu/elav. *J Exp Zool* **291**, 354-364, doi:10.1002/jez.1134 [pii] (2001).
227 Satou, Y. *et al.* Gene expression profiles in *Ciona intestinalis* tailbud embryos. *Development* **128**, 2893-2904 (2001).
228 Lisbin, M. J., Qiu, J. & White, K. The neuron-specific RNA-binding protein ELAV regulates neuroglian alternative splicing in neurons and binds directly to its pre-mRNA. *Genes Dev* **15**, 2546-2561, doi:10.1101/gad.903101 (2001).
229 Moroz, L. L. Phylogenomics meets neuroscience: how many times might complex brains have evolved? *Acta Biol Hung* **63 Suppl 2**, 3-19, doi:1362L96286258350 [pii] 10.1556/ABiol.63.2012.Suppl.2.1 (2012).
230 Valverde, R., Edwards, L. & Regan, L. Structure and function of KH domains. *FEBS J* **275**, 2712-2726, doi:EJB6411 [pii] 10.1111/j.1742-4658.2008.06411.x (2008).
231 Musco, G. *et al.* Three-dimensional structure and stability of the KH domain: molecular insights into the fragile X syndrome. *Cell* **85**, 237-245, doi:S0092-8674(00)81100-9 [pii] (1996).

232 Bassani, S. *et al.* The Neurobiology of X-Linked Intellectual Disability. *Neuroscientist*, doi:1073858413493972 [pii] 10.1177/1073858413493972 (2013).
233 Bhakar, A. L., Dolen, G. & Bear, M. F. The pathophysiology of fragile X (and what it teaches us about synapses). *Annu Rev Neurosci* **35**, 417-443, doi:10.1146/annurev-neuro-060909-153138 (2012).
234 Yano, M., Hayakawa-Yano, Y., Mele, A. & Darnell, R. B. Nova2 regulates neuronal migration through an RNA switch in disabled-1 signaling. *Neuron* **66**, 848-858, doi:10.1016/j.neuron.2010.05.007 S0896-6273(10)00375-2 [pii] (2010).
235 Zhang, C. *et al.* Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science* **329**, 439-443, doi:10.1126/science.1191150 science.1191150 [pii] (2010).
236 Winters, G. C. *et al.* Conserved expression patterns of Nanos in the ctenophore *Pleurobrachia bachei*: Implications for germ line specification in basal metazoans. *Integrative and Comparative Biology* **P3.77 Abstract** (2013).
237 Scheffer, G. L. *et al.* The drug resistance-related protein LRP is the human major vault protein. *Nat Med* **1**, 578-582 (1995).
238 Aouacheria, A. *et al.* Insights into early extracellular matrix evolution: spongin short chain collagen-related proteins are homologous to basement membrane type IV collagens and form a novel family widely distributed in invertebrates. *Mol Biol Evol* **23**, 2288-2302, doi:msl100 [pii] 10.1093/molbev/msl100 (2006).
239 Fowler, S. J. *et al.* Characterization of *Hydra* type IV collagen. Type IV collagen is essential for head regeneration and its expression is up-regulated upon exposure to glucose. *J Biol Chem* **275**, 39589-39599, doi:10.1074/jbc.M005871200 (2000).
240 Shimizu, H. *et al.* The extracellular matrix of *Hydra* is a porous sheet and contains type IV collagen. *Zoology (Jena)* **111**, 410-418, doi:10.1016/j.zool.2007.11.004 (2008).
241 Leys, S. P. & Riesgo, A. Epithelia, an evolutionary novelty of metazoans. *J Exp Zool B Mol Dev Evol* **318**, 438-447, doi:10.1002/jez.b.21442 (2012).
242 Churches, N., Kohn, A. B., Kokot, K. M., Swalla, B. J. & Moroz, L. L. Collagens in the ctenophore *Pleurobrachia bachei* :Remarkable expansion and diversity of genes controlling the extracellular matrix in basal metazoans. *Integrative and Comparative Biology* **P1.63 Abstract** (2013).
243 Milner, R. & Campbell, I. L. The integrin family of cell adhesion molecules has multiple functions within the CNS. *J Neurosci Res* **69**, 286-291, doi:10.1002/jnr.10321 (2002).
244 Sebe-Pedros, A., Roger, A. J., Lang, F. B., King, N. & Ruiz-Trillo, I. Ancient origin of the integrin-mediated adhesion and signaling machinery. *Proc Natl Acad Sci U S A* **107**, 10142-10147, doi:10.1073/pnas.1002257107 (2010).
245 Zhou, Y., Gunput, R. A. & Pasterkamp, R. J. Semaphorin signaling: progress made and promises ahead. *Trends Biochem Sci* **33**, 161-170, doi:10.1016/j.tibs.2008.01.006 (2008).
246 Nogi, T. *et al.* Structural basis for semaphorin signalling through the plexin receptor. *Nature* **467**, 1123-1127, doi:10.1038/nature09473 (2010).
247 Emes, R. D. & Grant, S. G. The human postsynaptic density shares conserved elements with proteomes of unicellular eukaryotes and prokaryotes. *Front Neurosci* **5**, 44, doi:10.3389/fnins.2011.00044 (2011).
248 Emes, R. D. & Grant, S. G. Evolution of synapse complexity and diversity. *Annu Rev Neurosci* **35**, 111-131, doi:10.1146/annurev-neuro-062111-150433 (2012).

249 Ryan, T. J., Emes, R. D., Grant, S. G. & Komiyama, N. H. Evolution of NMDA receptor cytoplasmic interaction domains: implications for organisation of synaptic signalling complexes. *BMC Neurosci* **9**, 6, doi:1471-2202-9-6 [pii] 10.1186/1471-2202-9-6 (2008).
250 Moroz, L. L. & Kohn, A. B. Parallel evolution of nitric oxide signaling: diversity of synthesis and memory pathways. *Front Biosci* **16**, 2008-2051, doi:3837 [pii] (2011).
251 Swore, J. J. *et al.* On the origins of glutamatergic signaling: Insights from the ctenophore genome (*Pleurobrachia bachei*). *Integrative and Comparative Biology* **33.6 Abstract** (2013).
252 Swore, J. J., KOHN, A. B., Citarella, M. R., Bobkova, Y. V. & Moroz, L. L. Molecular mapping of ctenophore neurons and glutamate signaling. *Integrative and Comparative Biology* **P2.132 Abstract** (2012).
253 Dingledine, R., Borges, K., Bowie, D. & Traynelis, S. F. The glutamate receptor ion channels. *Pharmacol Rev* **51**, 7-61 (1999).
254 Teng, H. *et al.* Evolutionary mode and functional divergence of vertebrate NMDA receptor subunit 2 genes. *PLoS One* **5**, e13342, doi:10.1371/journal.pone.0013342 (2010).
255 Martel, M. A. *et al.* The subtype of GluN2 C-terminal domain determines the response to excitotoxic insults. *Neuron* **74**, 543-556, doi:S0896-6273(12)00289-9 [pii] 10.1016/j.neuron.2012.03.021 (2012).
256 Diochot, S. *et al.* Black mamba venom peptides target acid-sensing ion channels to abolish pain. *Nature*, doi:nature11494 [pii] 10.1038/nature11494 (2012).
257 Sluka, K. A., Winter, O. C. & Wemmie, J. A. Acid-sensing ion channels: A new target for pain and CNS diseases. *Curr Opin Drug Discov Devel* **12**, 693-704 (2009).
258 Sherwood, T. W., Frey, E. N. & Askwith, C. C. Structure and activity of the acid-sensing ion channels. *Am J Physiol Cell Physiol* **303**, C699-710, doi:10.1152/ajpcell.00188.2012 ajpcell.00188.2012 [pii] (2012).
259 Agboh, K. C., Webb, T. E., Evans, R. J. & Ennion, S. J. Functional characterization of a P2X receptor from *Schistosoma mansoni*. *J Biol Chem* **279**, 41650-41657 (2004).
260 Bavan, S., Straub, V. A., Blaxter, M. L. & Ennion, S. J. A P2X receptor from the tardigrade species *Hypsibius dujardini* with fast kinetics and sensitivity to zinc and copper. *BMC Evol Biol* **9**, 17, doi:1471-2148-9-17 [pii] 10.1186/1471-2148-9-17 (2009).
261 Fountain, S. J. *et al.* An intracellular P2X receptor required for osmoregulation in *Dictyostelium discoideum*. *Nature* **448**, 200-203, doi:nature05926 [pii] 10.1038/nature05926 (2007).
262 Fountain, S. J. & Burnstock, G. An evolutionary history of P2X receptors. *Purinergic Signal* **5**, 269-272, doi:10.1007/s11302-008-9127-x (2009).
263 Wettschureck, N. & Offermanns, S. Mammalian G proteins and their cell type specific functions. *Physiol Rev* **85**, 1159-1204, doi:10.1152/physrev.00003.2005 (2005).
264 Bjarnadottir, T. K. *et al.* Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse. *Genomics* **88**, 263-273, doi:10.1016/j.ygeno.2006.04.001 (2006).
265 Lemmon, M. A. & Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **141**, 1117-1134, doi:10.1016/j.cell.2010.06.011 (2010).
266 Robinson, D. R., Wu, Y. M. & Lin, S. F. The protein tyrosine kinase family of the human genome. *Oncogene* **19**, 5548-5557, doi:10.1038/sj.onc.1203957 (2000).

267 Southey, B. R., Rodriguez-Zas, S. L. & Sweedler, J. V. Prediction of neuropeptide prohormone cleavages with application to RFamides. *Peptides* **27**, 1087-1098, doi:S0196-9781(06)00043-X [pii] 10.1016/j.peptides.2005.07.026 (2006).
268 Southey, B. R., Amare, A., Zimmerman, T. A., Rodriguez-Zas, S. L. & Sweedler, J. V. NeuroPred: a tool to predict cleavage sites in neuropeptide precursors and provide the masses of the resulting peptides. *Nucleic Acids Res* **34**, W267-272, doi:34/suppl_2/W267 [pii] 10.1093/nar/gkl161 (2006).
269 Amare, A. *et al.* Bridging neuropeptidomics and genomics with bioinformatics: Prediction of mammalian neuropeptide prohormone processing. *J Proteome Res* **5**, 1162-1167, doi:10.1021/pr0504541 (2006).
270 Duckert, P., Brunak, S. & Blom, N. Prediction of proprotein convertase cleavage sites. *Protein Eng Des Sel* **17**, 107-112, doi:10.1093/protein/gzh013 17/1/107 [pii] (2004).
271 Citarella, M. R., Kohn, A. B., Bobkova, E., Yu, F. & Moroz, L. L. Genome-wide characterization of signaling peptides across animal phyla and parallel evolution of neural systems. *Integrative and Comparative Biology* **P3.97 Abstract** (2011).
272 Citarella, M. R., Girado, D. O., Kohn, A. B. & Moroz, L. L. Global Discovery and Validation of Signaling Molecules in the Ctenophore. *Integrative and Comparative Biology* **P1.135 Abstract** (2012).
273 Abascal, F. & Zardoya, R. Evolutionary analyses of gap junction protein families. *Biochim Biophys Acta* **1828**, 4-14, doi:10.1016/j.bbamem.2012.02.007 S0005-2736(12)00058-2 [pii] (2013).
274 Connors, B. W. & Long, M. A. Electrical synapses in the mammalian brain. *Annu Rev Neurosci* **27**, 393-418, doi:10.1146/annurev.neuro.26.041002.131128 (2004).
275 Panchin, Y. V. Evolution of gap junction proteins--the pannexin alternative. *J Exp Biol* **208**, 1415-1419, doi:208/8/1415 [pii] 10.1242/jeb.01547 (2005).
276 Bao, L. *et al.* Innexins form two types of channels. *FEBS Lett* **581**, 5703-5708, doi:S0014-5793(07)01174-X [pii] 10.1016/j.febslet.2007.11.030 (2007).
277 Satterlie, R. A. & Case, J. F. Gap junctions suggest epithelial conduction within the comb plates of the ctenophore *Pleurobrachia bachei*. *Cell Tissue Res* **193**, 87-91 (1978).
278 Liebeskind, B. J., Hillis, D. M. & Zakon, H. H. Evolution of sodium channels predates the origin of nervous systems in animals. *Proc Natl Acad Sci U S A* **108**, 9154-9159, doi:1106363108 [pii] 10.1073/pnas.1106363108 (2011).
279 Shapiro, M. S. *et al.* Identification of subtypes of muscarinic receptors that regulate Ca2+ and K+ channel activity in sympathetic neurons. *Life Sci* **68**, 2481-2487 (2001).
280 Tompkins-Macdonald, G. J. *et al.* Expression of a poriferan potassium channel: insights into the evolution of ion channels in metazoans. *J Exp Biol* **212**, 761-767, doi:212/6/761 [pii] 10.1242/jeb.026971 (2009).
281 Gur Barzilai, M. *et al.* Convergent evolution of sodium ion selectivity in metazoan neuronal signaling. *Cell Rep* **2**, 242-248, doi:S2211-1247(12)00190-8 [pii] 10.1016/j.celrep.2012.06.016 (2012).
282 Cai, X. & Clapham, D. E. Evolutionary genomics reveals lineage-specific gene loss and rapid evolution of a sperm-specific ion channel complex: CatSpers and CatSperbeta. *PLoS One* **3**, e3569, doi:10.1371/journal.pone.0003569 (2008).
283 Brown, V. *et al.* Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. *Cell* **107**, 477-487, doi:S0092-8674(01)00568-2 [pii] (2001).

284 Ishibashi, K., Suzuki, M. & Imai, M. Molecular cloning of a novel form (two-repeat) protein related to voltage-gated sodium and calcium channels. *Biochem Biophys Res Commun* **270**, 370-376, doi:10.1006/bbrc.2000.2435 S0006-291X(00)92435-5 [pii] (2000).
285 Hooper, R., Churamani, D., Brailoiu, E., Taylor, C. W. & Patel, S. Membrane topology of NAADP-sensitive two-pore channels and their regulation by N-linked glycosylation. *J Biol Chem* **286**, 9141-9149, doi:M110.189985 [pii] 10.1074/jbc.M110.189985 (2011).
286 Brailoiu, E. *et al.* An ancestral deuterostome family of two-pore channels mediates nicotinic acid adenine dinucleotide phosphate-dependent calcium release from acidic organelles. *J Biol Chem* **285**, 2897-2901, doi:C109.081943 [pii] 10.1074/jbc.C109.081943 (2010).
287 Traynelis, S. F. *et al.* Glutamate receptor ion channels: structure, regulation, and function. *Pharmacol Rev* **62**, 405-496, doi:10.1124/pr.109.002451 62/3/405 [pii] (2010).
288 Omote, H., Miyaji, T., Juge, N. & Moriyama, Y. Vesicular neurotransmitter transporter: bioenergetics and regulation of glutamate transport. *Biochemistry* **50**, 5558-5565, doi:10.1021/bi200567k (2011).
289 El Mestikawy, S., Wallen-Mackenzie, A., Fortin, G. M., Descarries, L. & Trudeau, L. E. From glutamate co-release to vesicular synergy: vesicular glutamate transporters. *Nat Rev Neurosci* **12**, 204-216, doi:10.1038/nrn2969 nrn2969 [pii] (2011).
290 Palczewski, K. & Orban, T. From atomic structures to neuronal functions of G protein-coupled receptors. *Annu Rev Neurosci.*, ahead of print (2013).
291 Wemmie, J. A. *et al.* The acid-activated ion channel ASIC contributes to synaptic plasticity, learning, and memory. *Neuron* **34**, 463-477, doi:S089662730200661X [pii] (2002).
292 Krishtal, O. The ASICs: signaling molecules? Modulators? *Trends in neurosciences* **26**, 477-483, doi:S0166-2236(03)00210-8 [pii] 10.1016/S0166-2236(03)00210-8 (2003).
293 Bang, M. L. & Owczarek, S. A matter of balance: role of neurexin and neuroligin at the synapse. *Neurochem Res* **38**, 1174-1189, doi:10.1007/s11064-013-1029-9 (2013).
294 Papadopoulos, M. C. & Verkman, A. S. Aquaporin water channels in the nervous system. *Nat Rev Neurosci* **14**, 265-277, doi:10.1038/nrn3468 nrn3468 [pii] (2013).
295 Moroz, L. L. Genomic Bases for Independent Origins of Neurons and Complex Brains: New Insights from RNA-seq and genomic sequencing of basal metazoans, basal deuterostomes and molluscs. *Integrative and Comparative Neurology* **Abstracts**, http://www.sicb.org/meetings/2013/schedule/abstractdetails.php?id=2235 (2013).
296 Moroz, L. L. *et al.* The Genome of the Ctenophore *Pleurobrachia bachei*: Molecular Insights into Independent Origins of Nervous Systems *Integrative and Comparative Biology* Abstracts, http://sicb.org/meetings/2012/schedule/abstractdetails.php?id=2527 (2012).
297 Pennisi, E. Nervous system may have evolved twice. Society for Integrative and Comparative Biology Annual Meeting. *Science* **339**, 391, doi:doi:10.1126/science.339.6118.391-a 339/6118/391-a [pii] (2013).
298 Ryan, T. J. & Grant, S. G. The origin and evolution of synapses. *Nat Rev Neurosci* **10**, 701-712, doi:nrn2717 [pii] 10.1038/nrn2717 (2009).
299 Jekely, G. Origin and early evolution of neural circuits for the control of ciliary locomotion. *Proceedings. Biological sciences / The Royal Society* **278**, 914-922, doi:10.1098/rspb.2010.2027 (2011).

300 Miller, G. Origins. On the origin of the nervous system. *Science* **325**, 24-26, doi:325/5936/24 [pii] 10.1126/science.325_24 (2009).

301 Fodor, A., Kohn, A. B., Swalla, B. J. & Moroz, L. L. Quest for Muscle Specific Genes in *Pleurobrachia bachei* : Had mesoderm independently evolved in Ctenophores? *Integrative and Comparative Neurology (SICB)* Abstract, http://www.sicb.org/meetings/2013/schedule/abstractdetails.php?id=2455 (2013).

302 Swore, J. J. *et al.* On the Origins of Glutamatergic Signaling: Insights from the ctenophore genome (Pleurobrachia bachei). *Integrative and Comparative Neurology (SICB)* Abstract, http://www.sicb.org/meetings/2013/schedule/abstractdetails.php?id=2403 (2013).